Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference

Human Factors in high reliability industries

Edited by

Dick de Waard, Jürgen Sauer, Stefan Röttger, Annette Kluge, Dietrich Manzey, Clemens Weikert, Antonella Toffetti, Rebecca Wiczorek, Karel Brookhuis, and Jettie Hoonhout

ISSN 2333-4959 (online)

Please refer to contributions as follows:

[Authors] (2015), [Title]. In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference (pp. pagenumbers). Downloaded from http://hfes-europe.org (ISSN 2333-4959)

Available as open source download

Published by HFES

Contents

HUMAN MACHINE INTERACTION

Advantages of Magnetic Mice over Trackballs as input devices on moving platforms
Stefan Röttger, Saskia Vetter, & Sören Ollhoff

Investigation of human behaviour in pushing and pulling tasks for direct manipulation of a collaborative robot
Jonas Schmidtler, Christina Harbauer, & Klaus Bengler

Validation of a Telephone Manager for stressful driving situations
Linda Köhler, Klaus Bengler, Christian Mergl, Kathrin Maier, & Martin Wimmer

Anger and bother experience when driving with a traffic light assistant: A multi-driver simulator study
Lena Rittger, Dominik Muehlbacher, Christian Maag, & Andrea Kiesel

Olfaction influences affect and cognitive-motoric performance: Evidence for the negative impact of unpleasant odours.
Stefan Brandenburg, Anna K. Trapp, & Nils Backhaus

AUTOMATION

The more the better? The impact of number of stages of likelihood alarm systems on human performance.
Magali Balaud & Dietrich Manzey

The predictive quality of retentivity for skill acquisition and retention in a simulated process control task
Barbara Frank & Annette Kluge

Implementing dynamic changes in automation support using ocular-based metrics of mental workload: a laboratory study
Serena Proietti Colonna, Claudio Capobianco, Simon Mastrangelo, & Francesco Di Nocera

AVIATION

Event expectancy and inattentional blindness in advanced helmet-mounted display symbology
Patrizia Knabl, Sven Schmerwitz, & Johannes Ernst

A novel Human Machine Interaction (HMI) design/evaluation approach supporting the advancement of improved automation concepts to enhance flight safety
Joan Cahill & Tiziana C. Callari

Option generation in simulated conflict scenarios in approach Air Traffic Control
Jan Kraemer & Heinz-Martin Süß

The operational potential of an In-Flight Weather Awareness System: an explorative pilot-in-the-loop simulation
Simone Rozzi, Stefano Bonelli, Ana Ferreira, Linda Napoletano, & Loic Bécouarn

Innovative multi-sensor device deployment for fighter pilots activity study in a highly realistic Rafale simulator
Julie Lassalle, Philippe Rauffet, Baptiste Leroy, Laurent Guillet, Christine Chauvin
& Gilles Coppin

TRANSPORTATION

Is simulation (not) enough? Results of a validation study of an autonomous emergency braking system on a test track and in a static driving simulator.
Martin Jentsch & Angelika C. Bullinger

Success factors for navigational assistance: a complementary ship-shore perspective.
Linda de Vries

Can weak-resilience-signals (WRS) reveal obstacles compromising (rail-)system resilience?
Willy Siegel & Jan Maarten Schraagen

Introducing electric vehicle-based mobility solutions – impact of user expectations to long and short term usage
André Dettmann, Dorothea Langer, Angelika C. Bullinger, & Josef F. Krems

Are globality and locality related to driver's hazard perception abilities?
Shani Avnieli-Bachar, Avinoam Borowsky, Yisrael Parmet, Hagai Tapiro, & Tal Oron-Gilad

TRAINING

How to improve training programs for the management of complex and unforeseen situations?
Marie-Pierre Fornette, Françoise Darses, & Marthe Bourgy
The Expanded Cognitive Task Load Index (NASA-TLX) applied to Team Decision-Making in Emergency Preparedness Simulation
Denis A. Coelho, João N. O. Filipe, Mário Simões-Marques, & Isabel L. Nunes

MEDICINE

Evaluation of Crew Resource Management Interventions for Doctors-on-call
Vera Hagemann, Annette Kluge & Clemens Kehren

TESTING & EVALUATION

Can we remove the human factor from usability research to save time and money?
Andreas Espinoza & Johan Gretland
Influence of head mounted display hardware on performance and strain.
Matthias Wille & Sascha Wischniewski
Title
Jettie Hoonhout

Advantages of Magnetic Mice over Trackballs as input devices on moving platforms

Stefan Röttger[1], Saskia Vetter[1], & Sören Ollhoff[2]
[1]German Naval Medical Institute
[2]Federal Office of Bundeswehr Equipment, Information Technology and In-Service Support
Germany

Abstract

Although ergonomic studies show that cursor control with a computer mouse is faster and sometimes more accurate than cursor control with a trackball, trackballs are the standard input device for cursor movements on many moving platforms such as airplanes and ships. One reason for this is that trackballs can be fixed to the workstation, which prevents involuntary cursor movements that could otherwise be induced by movements of the platform. In this study, standard trackballs and computer mice with magnetic adhesion to the mouse pad were evaluated by 18 sailors of the German Navy after 26 days of computer operation on their moving ship. Results show that users of magnetic mice performed better and showed less muscular fatigue than trackball users. Thus, magnetic mice should be considered as the standard input device on moving platforms.

Introduction

Although the standard input device for cursor control in the operation of most computer systems is the computer mouse, trackballs are commonly used for cursor control on moving platforms such as ships or airplanes. There are two reasons for this preference of trackballs: first, on many moving platforms, there is only limited space to accommodate the human-computer-interface and less space is required for the operation of a trackball. Second, trackballs can be fixed to the workplace, which is intended to prevent motion-induced shifts of the device and the cursor on the computer screen.

Ergonomic research has found that compared to mouse use, trackball use can be associated with a number of disadvantages. Studies of user performance in fixed laboratory settings show that computer mice allow for a faster and more precise cursor control than trackballs (Grandt et al., 2004; Isokoski et al., 2007). Similar results were obtained in an experiment with participants experiencing simulated ship movements while performing a Fitts task. Trackball-controlled cursor movements to a target location were as accurate as mouse-controlled cursor movements, but on average 500 ms slower (Lin et al., 2010). Results on muscular strain associated with mouse and trackball use are rather inconclusive. While trackball use during a five-minute period of office work was found to cause less muscular activity in shoulder

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

and neck, it led to a higher wrist extension than mouse use. Neither of these differences was reflected in the subjective strain ratings of the study participants (Karlqvist et al., 1999).

The studies referred to above were mostly conducted in stationary environments and with rather short periods of work. The objective of the present investigation was to study performance and strain differences between mouse use and trackball use on a seagoing platform and over extended periods of time. The computer mice used in this study were secured against motion-induced shifts by magnetic adhesion to the mouse pad. This results in a certain resistance that must be overcome when starting to move the mouse. Another purpose of this study was to find out whether the computer mice thus modified would show the same advantages over trackballs as the standard devices used in previous studies.

Methods

A sample of 18 male sailors of the German Navy participated in this study. They performed their usual tasks with a computer system in the Combat Information Centre (CIC) of a German frigate (for an example of typical workstations in a CIC see figure 1).

Tasks included the radar-based detection and classification of airplanes and vessels, acquisition of potential threads, thread engagement and weapon control. Type of input device was manipulated in a between subjects design. Ten participants used a recessed trackball and eight participants used an optical mouse as input device. Inside their housing, the mice were equipped with neodym magnets that provided adhesion to ferromagnetic mouse pads. Special care was taken to keep the magnetic adhesion and thus the necessary force to overcome the adhesion when moving the mouse as low as possible. Participants tested their input device for a period of 26 days during transit voyages and a weapon exercise. The mean duration of consecutive computer operation was four to six hours each day. Wave heights during the trial period were between 0.5 and 4 metres.

After the end of the trial period, participants gave their subjective evaluation of the input device on a seven-point rating scale with the questionnaire from ISO 9241-420, appendix D.1. This questionnaire contains items regarding the performance in cursor control (speed, accuracy, smoothness of cursor movements), the difficulty of operating the device (force, effort), and fatigue of fingers, wrist, arm, shoulder and neck. Higher ratings in this questionnaire indicate a better evaluation. Two additional scales of the questionnaire with summary ratings (overall satisfaction and usability) were not considered in the analysis because they contain no additional information beyond the specific items on performance, difficulty and muscular fatigue.

Figure 1. Typical workstations in a Combat Information Centre of a German frigate. Note the recessed trackball at the bottom of the picture. © Bundeswehr.

Ratings of mouse and trackball users were compared with t-tests for independent samples. Due to the multiple testing, a Šidàk-correction (Abdi, 2007) was applied and the test-wise alpha level was set to .0051 in order to keep the family-wise alpha level at 0.05.

Results

Means, standard deviations and test statistics of all items are displayed in table 1. Results regarding performance, difficulty and muscular fatigue are summarized below the table. In the box plots used for graphical data representation, horizontal bars indicate the median of the distribution. Boxes cover the central 50% of the data range and vertical lines cover observed values of up to 1.5 times the central data range. Individual values beyond that point are represented by dots.

Performance

The distribution of the performance ratings is illustrated in figure 2. The magnetic mouse received significantly better mean ratings on all performance items of the questionnaire, i.e. speed (6.4 vs. 2.5, $p<.001$), accuracy (6.3 vs. 4.2, $p<.001$) and smoothness of movements (5.8 vs. 3.4, $p<.001$).

Difficulty

The data of the difficulty ratings are depicted in figure 3. For a more intuitive comprehension of the plot, values were reflected to have higher levels of force and effort indicated by higher values. Mouse users reported significantly more

comfortable levels of force required in the use of their input device (5.8 vs. 3.4, p=.003). The average effort ratings did not differ significantly (p>.0051).

Table 1: Descriptive and inferential statistics of mean ratings of trackball and magnetic mouse.

	Magnetic Mouse		**Trackball**		**t-test**		
Item	M	SD	M	SD	t	df	p
1. Force	5.8	1.3	3.4	1.5	3.58	15.9	.0025
2. Smoothness	5.8	0.9	3.4	1.1	5.08	16.0	.0001
3. Effort	6.0	1.8	3.2	1.9	3.20	15.6	.0057
4. Accuracy	6.3	0.7	4.2	1.2	4.44	14.7	.0005
5. Speed	6.4	0.7	2.5	1.7	6.43	12.8	>.0001
6. Satisfaction	6.5	0.8	2.3	1.4	8.05	14.2	>.0001
7. Overall usability	6.4	0.7	3.5	1.6	4.92	13.1	.0003
8. Fatigue of finger	6.5	0.9	2.9	1.3	6.89	15.9	>.0001
9. Fatigue of wrist	6.4	0.7	2.9	1.7	5.91	13.0	>.0001
10. Fatigue of arm	6.0	0.9	3.6	1.7	3.79	14.3	.0019
11. Fatigue of shoulder	6.0	1.2	3.8	1.9	2.96	15.2	.0096
12. Fatigue of neck	5.9	1.6	3.3	2.1	2.91	16.0	.0102

Notes. M: Mean rating on a scale from 1-7. SD: standard deviation. t: test-statisticof t-test. df: degrees of freedom, corrected for inequalities of variances. p: significance. The table contains questionnaire data in their original form, with higher values consistently indicating more positive evaluations (e.g. less fatigue, more accuracy).

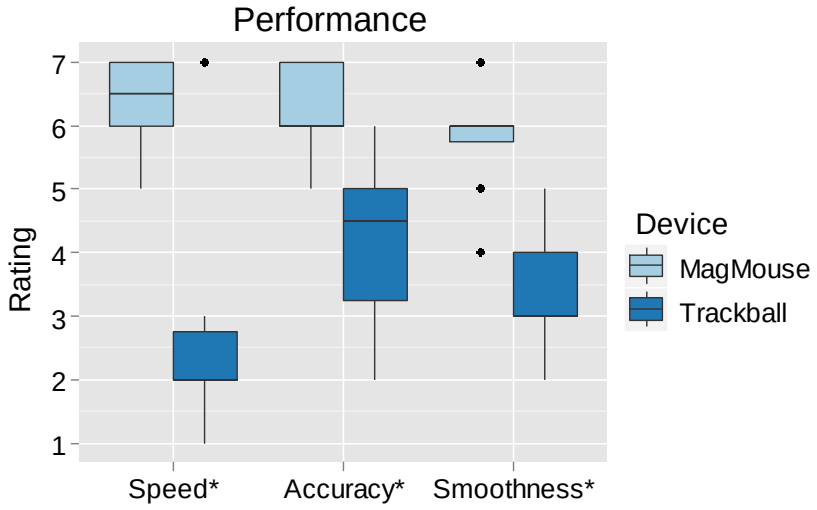


Figure 2. Boxplot of performance ratings of mouse users and trackball users. Significant differences (p<.0051) are marked with an asterisk. Higher values denote better performance

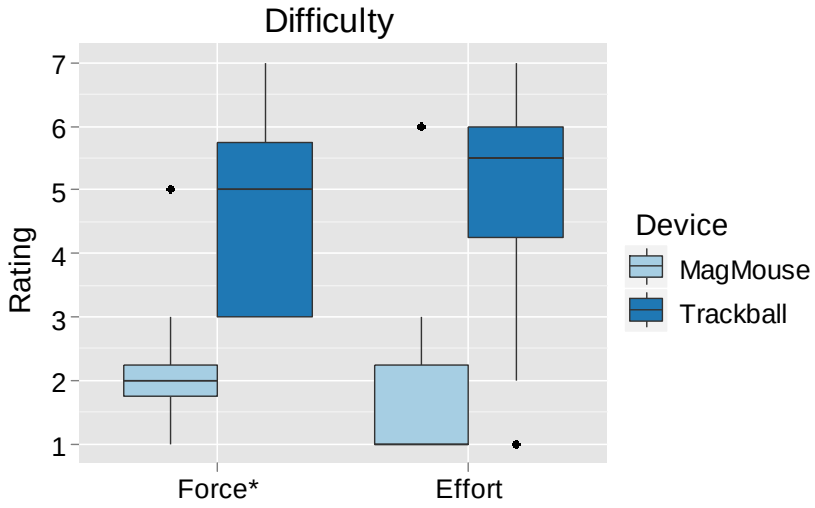


Figure 3. Boxplot of difficulty ratings of mouse users and trackball users. Data were mirrored for graphical depiction, higher values indicate higher levels of force and effort. Significant differences (p<.0051) are marked with an asterisk.

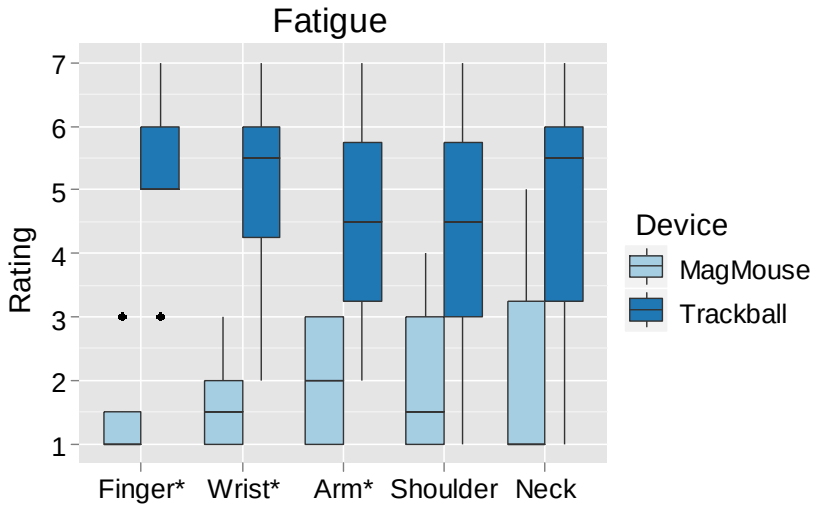


Figure 4. Boxplot of fatigue ratings of mouse users and trackball users. Data were mirrored for graphical depiction, higher values indicate higher levels of fatigue. Significant differences (p<.0051) are marked with an asterisk.

Fatigue

An overview over the reflected fatigue ratings of mouse and trackball users can be found in figure 4. Significantly less fatigue, as indicated by better and thus higher

ratings in the questionnaire, were found for fingers (6.5 vs. 2.9, $p<.001$), wrist (6.4 vs. 2.9, $p<.001$) and arm (6.0 vs. 3.6, $p=.002$) of mouse users. No significant difference was found for experienced fatigue in shoulder and neck ($p>.0051$).

Conclusion

This paper presented a field study on the consequences of cursor control with trackballs and magnetic mice. Compared to laboratory investigations of this topic, we could exercise rather little experimental control.

Although the tasks accomplished with magnetic mice and the trackballs were reported to be similarly demanding, they were not identical. And we could reasonably assume, but not assure that the participants of the mouse group and the trackball group had the same level of proficiency in computer operation. Thus, the internal validity of our study is lower than that of laboratory investigations.

However, our study was conducted to add results with a higher external validity to the literature on mouse use and trackball use. To this end, the investigation was carried out on a moving ship, with the actual tasks of operators, over an extended period of 26 days with 4-6 hours of consecutive computer operation each day. Under these circumstances, the previously reported performance advantages of mice over trackballs (Grandt et al., 2003; Isokoski et al., 2007; Lin et al., 2010) were replicated with mice that were magnetically secured against involuntary movements. Despite the necessity to overcome the magnetic adhesion of the mice when starting to move them, the data show that use of a magnetic mouse still leads to less muscular strain of the operators than the use of a trackball.

Interestingly, the differences in experienced muscular strain found in our study did not occur in the study of Karlqvist et al. (1999), which is most probably owed to the much shorter task duration of only 15 minutes in that study. Another noteworthy pattern of results is that the strain difference between mouse and trackball becomes the smaller the more distal the rated body part is from the input device. Based on informal observations, we assume that the higher strain of fingers and wrist is caused by the fact that these parts of the body have to move more often and to cover longer distances to produce the same cursor movement on the screen with a trackball as compared to a mouse.

To sum up, it can be concluded that the use of magnetic mice instead of trackballs is beneficial for operators' performance, for their health and thus for their long-term work capability. Designers of computer workstations for moving platforms should consider magnetic mice as the standard input device for cursor control and should be aware that the advantage of trackballs in modest space requirement trades off with disadvantages in operator strain and performance.

References

Abdi, H. (2007). The Bonferonni and Šidák Corrections for Multiple Comparisons. In N. Salkind (Ed.) *Encyclopedia of Measurement and Statistics* (pp. 103-107). Sage, Thousand Oaks (CA).

DIN EN ISO 9241-420. Ergonomics of human-system interaction – Part 420: Selection of physical input devices. Berlin, Beuth-Verlag.

Grandt, M., Pfendler, C., & Mooshage, O. (2003). Empirical Comparison of Five Input Devices for Anti-Air Warfare Operators. *Proceedings of the 8th International Command and Control Research and Technology Symposium*, retrieved 30.04.2013 from URL: www.dodccrp.org/events/8th_ICCRTS/pdf/035.pdf.

Isokoski, P., Raisamo, R., Martin, B., & Evreinov, G. (2007). User performance with trackball-mice. *Interacting with Computers, 19*, 407-427.

Karlqvist, L., Bernmark, E., Ekenvall, L., Hagberg, M., Isaksson, A., & Rostö, T. (1999). Computer mouse and track-ball operation: Similarities and differences in posture, muscular load and perceived exertion. *International Journal of Industrial Ergonomics, 23*, 157 - 169.

Lin, C.J., Liu, C.N., Chao, C.J., & Chen, H.J. (2010). The performance of computer input devices in a vibration environment. *Ergonomics, 53*, 478-490.

Investigation of human behaviour in pushing and pulling tasks for direct manipulation of a collaborative robot

Jonas Schmidtler, Christina Harbauer, & Klaus Bengler
Institute of Ergonomics, Technische Universität München
Germany

Abstract

This study is concerned with the human behaviour while pushing and pulling a trolley to get information about the characteristics of the human part in a physical human-robot interaction. The trolley was laden with three different weights and three different object sizes that should separate the connection between estimated weight and exerted force. The participants had to push and pull the trolley over a given path, similar to a real production scenario, e.g. in automotive assembly lines. Twenty-two people participated and were monitored by a VICON motion tracking system. The applied forces were gathered independently on each handle in three coordinates via a Kistler hand force measuring system. Results show that humans accelerate faster (jolt), higher (a), and get to higher velocities (v) when a certain amount of force is needed. Consequently enough feedback has to be implemented in novel collaborative assistant systems.

Introduction

Motivation – Why do we need Human-Robot Collaboration (HRC)?

The production environment faces decisive trends nowadays that cause a rethinking of classical production schemes. The upcoming customization of production (Fogliatto, da Silveira & Borenstein, 2012, Da Silveira, Borenstein & Fogliatto, 2001) stands contradictory to the continuing trend of mechanization and automation of work systems (Schlick, 2009). Mass customization is characterized by a customer orientation that causes decreasing lot sizes and increasing variety that have to be managed by flexible production systems. Present automation cannot fulfil the required flexibility and the presence of the human worker will still be necessary. In the assembly context Lotter and Wiendahl (2006) postulate the cost-optimum at a system called *hybrid assembly system* where manual tasks, operated by human workers, are combined with automatic contents.

Especially in the assembly area as the last link in the value chain and still the most employee-intensive area of the production, the designer of new solutions should always take the human with his needs and capabilities in consideration. Human

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

abilities like fast perception and processing of various information or flexible adaption and improvisation can be the key success factor for handling tasks. If it is possible to bring these benefits together with advantages of automation such as precision, strength, and reproducibility of robots, many problems could be solved at a time. Intelligent Assist Devices (IADs), also called Cobots, are able to bring these facts together and combine the characteristics of industrial robots and manual operated handling devices that are already common in automotive assembly lines (Akella et al., 1999).

Cobots – collaborative robots for handling tasks

The word Cobot (collaborative robot) was introduced by Michael Peshkin and J. Edward Colgate, associate professors of Mechanical Engineering at the Northwestern University, USA. Based on Peshkin and Colgate (1999) Cobots are meant to be used in direct interaction with a human worker, handling a payload together in a designated collaboration area (DIN EN ISO 10218-2). The goal is to close the gap between the stated limits and combine the respective advantages of each other: easy operation and low cost of the manipulators on the one hand and the precision, programmability and path guidance of an industrial robot on the other. Physical interaction with a Cobot enables strength amplification, inertia masking (starting, stopping, and turning forces) and guidance via virtual surfaces (walls, paths) (Colgate, Peshkin, & Klostermeyer, 2003). They are able to support the human not only in a physical but also in a cognitive way. These assistance systems can be used to facilitate handling tasks while increasing the efficiency of the process itself. Unlike industrial robots they are not separated from people because of safety reasons. They are able to improve ergonomic working conditions, product quality, and productivity (Peshkin & Colgate, 1999).

The possibility to implement virtual surfaces in the handling process is one crucial advantage of the new technology (figure 1). For clarification virtual surfaces can be described by the analogy to the role of a straightedge in drafting (Peshkin & Colgate, 1999). The virtual surfaces as well as the straightedge provide physical guidance along a defined shape path but it leaves the decision to the operator to use it (push payload up against) or not (pull away and guide payload by the worker himself). In this way an important ergonomic improvement arises. By supporting lateral and stabilizing forces on a payload, stress to the muscles of the upper body and whole back can be minimized. The virtual walls or paths could additionally be used for obstacle avoidance like virtual fences that surround and protect objects in the workspace. Furthermore through virtual guidance it is possible to increase the efficiency by precise and quick assembly processes while the cognitive workload on the human operator is getting reduced.

The second main advantage of a Cobot is to support the human operator in the handling task by reducing the required forces (figure 1). With power assistance (compensation of frictional and acceleration/deceleration forces) and force amplification (compensation of inertial, gravitational and frictional forces) the Cobot assists the human worker in handling large unhandy objects (Akella et al., 1999). In this way not only the human strength is amplified also inertia forces (starting,

stopping and turning forces) that act on the human body are getting masked and musculoskeletal disorders (MSDs) can be prevented.

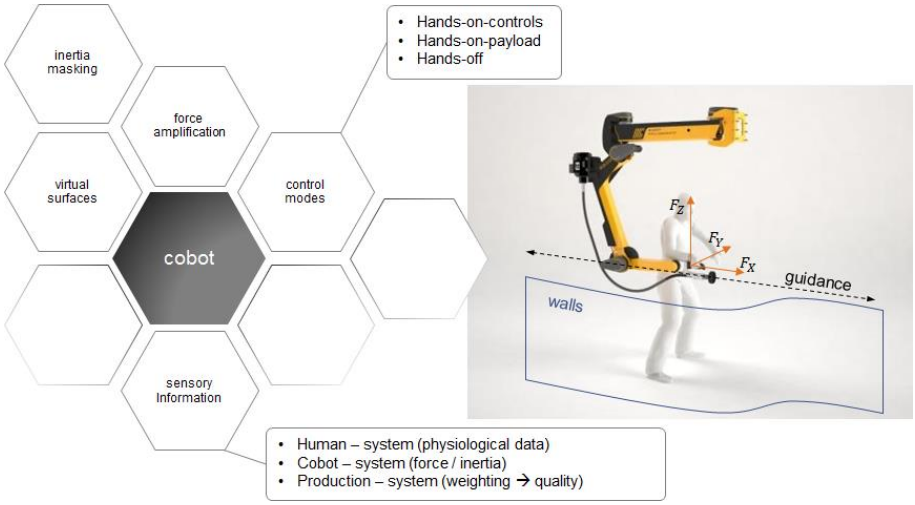


Figure 1. Capabilities of a new class of material handling devices; Cobot example: RB3D (2014)

Basically three modes of operation are conceivable with a Cobot (Robotic Industries Association, 2002): *Hands-on-controls mode* when the operator is in physical control from a designated control interface (e.g. handles), *hands-on-payload* when the powered motion is in response to forces applied directly to the payload and *hands-off control mode* where the motion follows a pre-determined path and is not in response to forces applied by the operator. A fourth control mode might be a hybrid form of *hands-on-controls mode* and *hands-on-payload* where the user can manipulate the position of the payload relatively to the Cobot. This scheme explains the semi-automatic abilities of a Cobot system. While in hands-on-control and hands-on-payload mode the user executes mainly manual tasks, supported by the automation, the Cobot is able to act autonomously in hands-off control mode. Functions like *return-to-home* or *bring-the-next-part* can reduce operation time and the process gains flexibility and efficiency. Besides these functionalities Cobots also provide benefits by offering an interface to sensors for special purposes ,e.g. weighing parts or tracking moving assembly lines, and provide plant information systems, for error-proofing and data logging (Colgate et al., 2003).

Research topic

As said before in *hands-on-controls/hands-on-payload* mode the operator is in direct contact with the Cobot/payload and experiences a reaction force. Simultaneously sensing the intention of the human operator and how much feedback he gets is of central importance. According to that the main research topic in the field of cobotics for the Institute of Ergonomics is to examine the human characteristics while performing pushing and pulling tasks with and without power assisted and force amplified systems in detail. On the one hand the haptic feedback should be designed

like it is most natural for the human operator, ideally as if the worker is performing the task fully manual (Colgate et al., 2003) and on the other hand the handling task must not demand too little from the worker. Because acceptance of the new systems depends directly on the sensitivity, intuitiveness, and transparency of the haptic interface and its interpretation, it is crucial to understand how the human reacts while pushing/pulling a Cobot and what they actually sense. Before the design and implementation of a novel Cobot control system preliminary tests have to be used to investigate the human in pushing/pulling tasks.

Method

Motivation & hypotheses

The main goal of this study was knowledge-acquisition on intuitive kinaesthetic collaboration in pushing and pulling tasks. Studying the interaction of a human with a non-powered trolley should provide a database to design the direct physical Human-Robot Interaction of a novel Cobot system. The conducted study should give insight whether it is possible to develop a model for the human behaviour in pushing and pulling tasks and which performance parameters can be used for this purpose. Research has been already done in the field of haptic interaction. Groten (2011) for example measured mutual haptic interaction in her dissertation about Human Dyads – a method to investigate and optimize haptic interaction – in task performance, the physical effort, and efficiency (combination of the first two measurements). Since these factors cannot be easily implemented in a real-time system, it became necessary to begin at a former step. So two questions arise in the context of a new collaborative assistance system, which should be answered before further studies can be conducted.

Does the size of the handled object influence the operator's intention of how much force he should apply to manoeuvre the payload? Hence the first hypothesis reads as follows:

H1: There is a relationship between object size and expected weight in pushing / pulling tasks.

The second main question is, if there is a mismatch between expected and experienced weight of the payload, are there any variances in the movement parameters (velocity, acceleration, and jolt) while pushing and pulling a trolley? Hence the second hypothesis reads as follows:

H2: The weight-size mismatch has a significant influence on velocity (v), acceleration (a), and jolt (j).

Framework conditions of the study

The trolley, the laden weights, and the visual objects
The study included a trolley laden with three different, for the participants invisible, weights (0, 20 & 60 kg) and three different visible objects on top of the trolley (figure 2, left). The trolley is comparable with a serving cart for common tasks like

commissioning. It holds two platforms which were used to carry the payload on the lower one and the object sizes on top. The four-wheeled trolley was modified as shown in figure 2. The whole space below the upper platform was covered by black cardboard to hide the laden weights form the participants' eyes. Blackened aluminium profiles were mounted on the cart to allow an adaptive handle height, distance, and orientation. In this way a comparable application of force for any anthropometric requirements of each participant could be ensured. The floor of the experimental room was made of PVC and manoeuvring the cart was smooth and without any irregularities.

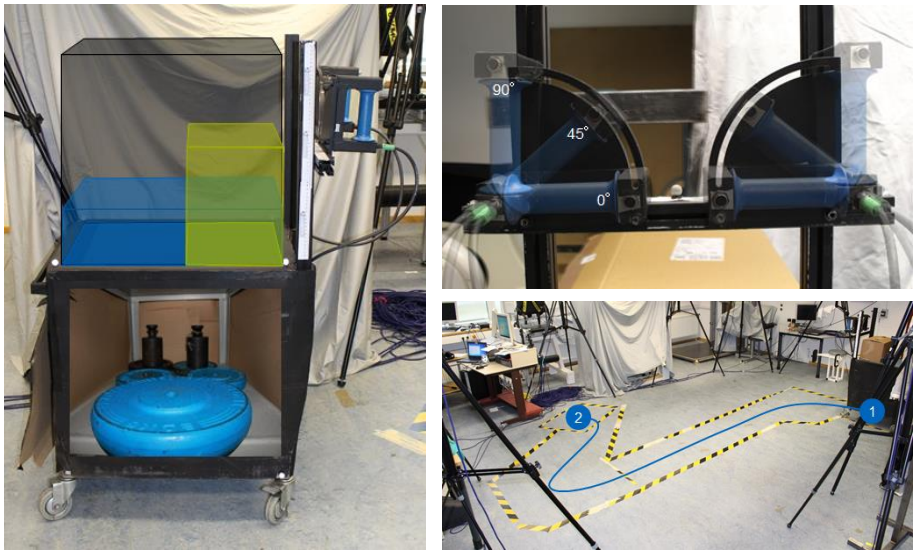


Figure 2. (left) Trolley laden with three different object sizes on top and three different weights hidden in the "belly"; (right) 90° angle of the handles and the path beginning at position 1 and ending at position 2.

Vicon Motion Tracking

The motion of operator and trolley was tracked by ten Vicon T160 cameras which were placed around the experimental area. They capture at 120fps with 16 megapixel (4704 x 3456). Vicon Nexus 1.8.2 had been used for processing the motion data and transferring it to .cvs format. The system provides Cartesian coordinates of each marker – in x, y, and z – related to an initial coordinate system. (Bortot et al., 2010) The information about the position of each marker for each frame were edited with a MATLAB script. By nominalization of the x-y vector and numerical derivation, a five-point stencil in one dimension, the first derivate of position, velocity, and acceleration had been made.

$$f'(x) \approx \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h}$$

In this way it was possible to get information about velocity (v), acceleration (a), and jolt (j) for any recorded frame.

Kistler Hand Force Measuring System
The Hand Force Measuring System for Ergonomics, Biomechanics and Occupational Health & Safety (Type 9809A) from Kistler (2014) was attached to the trolley (figure 2). It records the three orthogonal force components at 50 Hz with a piezoelectric multicomponent system. The system logs related to a Cartesian coordinate system defined at the trolley's front left wheel.

Subjective ratings
To measure the subjective expected as well as the experienced strain an in-house developed survey was applied. The participants were asked to rank their opinion in a scale from no strain (1) till very high strain (5).

Experimental design
The study was conducted in an experimental room at the Institute of Ergonomics. A five metre long given path, similar to a real production scenario in automotive assembly lines, were marked on the floor (figure 2, right). The participants had to start at point 1 pull the trolley back, turn it right, push it all the way to the end of the straight line, again turn it right, and push it to the position 2. Marker for the motion tracking system were positioned on the operator and the cart (figure 3). Each participant was marked with nine markers on hand, elbow, shoulder, neck, lower, and upper chest. The trolley was marked with seven markers on the top platform, side, and between the handles.

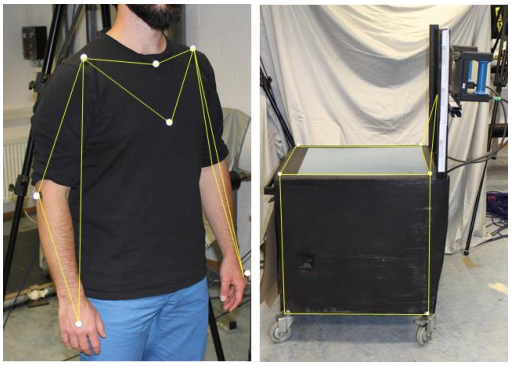


Figure 3. Marker position on the operator and trolley. Markers were placed on the upper chest of the participant and in the middle of the two handles on the trolley.

Procedure

At the beginning of the experimental session (*preparation phase*), all participants were asked to state demographic data like gender, age, and sportiness. In this study sportiness refers to the number of days within a seven-day week in which sport actively is performed (exercise, swimming etc.). Anthropometric data of each participant were gathered for body weight and height, solar plexus height, shoulder height and width, forearm length, upper arm length, handle height, and handle distance. General questions like the expected strain in a panoramic sunroof in an automobile assembly were asked to prime the participants for the simulated situation. In the next phase (*expectations*) the participants had to push / pull the

trolley in three stages. These three conditions differed in the handled payload (0, 20, 60 kg) and three object sizes (1 – small, 2 – medium, 3 – large). Every participant started with the 20kg-medium condition and followed either with 60kg–small (group 1) or 0kg-large (group 2) followed by the other condition as third condition. Before and after every condition the participants were asked about their subjective strain (expected respectively experienced). The following phase (*handle orientation*) was dedicated to investigate three handle orientations (0°, 45°, 90° angle) relating to the three weights mentioned above (nine stages). To qualify the observed forces, the maximum forces of each participant in 15 states were measured (*maximum force measurement*). The last two phases are not included in this paper.

Participants

Twenty-two healthy volunteers participated in this study (13 men, 9 women). The participants were between 21 and 32 years of age (SD = 2.6). No participants reported to suffer from any motoric impairment. 16 of them indicated to regularly do sports (M = 3.07 days / week), thereof 11 endurance and 5 weight training. Table 1 depicts the anthropometric measurements of the participants interrelate to percentile scores provided in the SizeGermany data (Seidl, Trieb, & Wirsching, 2008). Body weight and handle distance of the sample cover almost the whole typology whereas body height and handle height only take into account about half of the possible measurements. All participant data were normally distributed (Kolmogorov-Smirnov-test VALUE= , NS).

Table 1. Anthropometric measurements of the participants related to the percentile values of the SizeGermany data (Seidl, Trieb, & Wirsching, 2008)

	M	*SD*	*relation to SizeGermany*	
body weight	74.80 kg	14.44 kg	3p woman	98p man
body height	176.25 cm	7.85 cm	51p woman	95p man
handle height	109.45 cm	5.52 cm	47p woman	98p man
handle distance	37.27 cm	3.55 cm	5p woman	95p man

Results

Statistical analysis

One-way repeated measures ANOVA and a paired-samples t-test were conducted for statistical analysis. Degrees of freedom were corrected using the Greenhouse-Geisser correction factor if the criterion of sphericity was not met. For all analyses, the significance level was set to 0.05. Analyses of the force and movement values revealed that the median should be taken into account. Basis for this decision is the advantage of the median that this measure is insensitive to outliers.

Forces

Before the three mentioned weight-size-mismatches could be investigated it had to be clarified if the loadings create three significant different conditions. Table 2 illustrates the arithmetically averaged median of the forces for the three states 20kg-

medium, 60kg-small, and 0kg-large. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2) = 6.068$, $p = .048$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .793$). The results show that there was a significant effect of the load condition on mean applied forces $F(1.585, 33.289) = 55.502$, $p < .001$. These results suggest that the three different weight-size-mismatches create three different experimental conditions. Post-hoc Bonferroni comparisons indicated that all applied forces were significantly different from each other, $p \leq .001$.

Table 2. Median forces for the three conditions 20kg-medium, 60kg-small, and 0kg-large arithmetically averaged

	20kg-medium	*60kg-small*	*0kg-large*
M	41.48 N	60.78 N	33.24 N
SD	8.53 N	17.52 N	9.80 N

Expected and experienced strain

Since the three loadings can be seen as three different experimental conditions it was object of contemplation if there is a correlation between object size and expected strain. Each participant had to assess the awaited strain just by looking at the laden trolley with the object placed on its upper platform. Figure 6 shows the mean of the subjective expected strain for each condition and divided in the two mentioned groups. Mauchly's test indicated that the assumption of sphericity had been not violated, $\chi^2(2) = 1.173$, $p = .556$, therefore sphericity can be assumed. The results show that there was a significant effect of the object size on mean expected strain $F(2,42) = 19.958$, $p < .001$. These results suggest that the object size has an influence on the estimation of strain. Post-hoc Bonferroni comparisons indicated that the conditions 20kg-medium and 60kg-small were significantly different from 0kg-large, $p \leq .001$. The two conditions among themselves were not significantly different, $p = .150$.

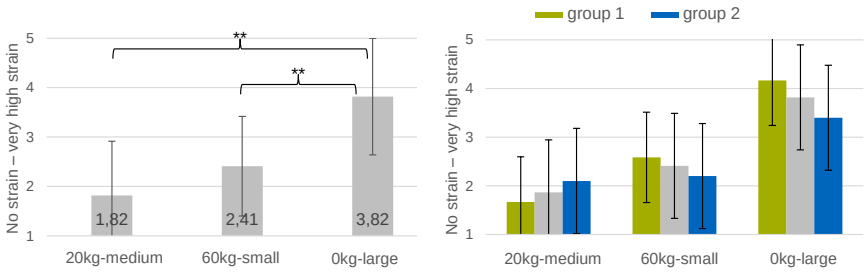


Figure 6. Mean of subjective expected strain depending on laden object size

With this in mind the difference between expected and experienced strain was of interest. Figure 7 depicts the mean statements of the participants before and after pushing / pulling the trolley over the trail. A paired-samples t-test was conducted to compare the statements before and after the task. The results show that there are

significant differences in the scores for 60kg-small expected (M = 2.41, SD = 1.01) and experienced (M = 3.73, SD = 0.94), t(21) = -5.11, p < .001, and 0kg-large expected (M = 3.82, SD = 1.18) and experienced (M = 1.07, SD = 0.32), t(21) = 10.16, p < .001. The two scores in the 20kg-medium condition (expected: M = 1.82, SD = 1.10; experienced: M = 2.10, SD = 0.81) were not significant different, t(21) = -1.19, p > .05. This leads to the assumption that the participants assume the strain of the task because of the object size.

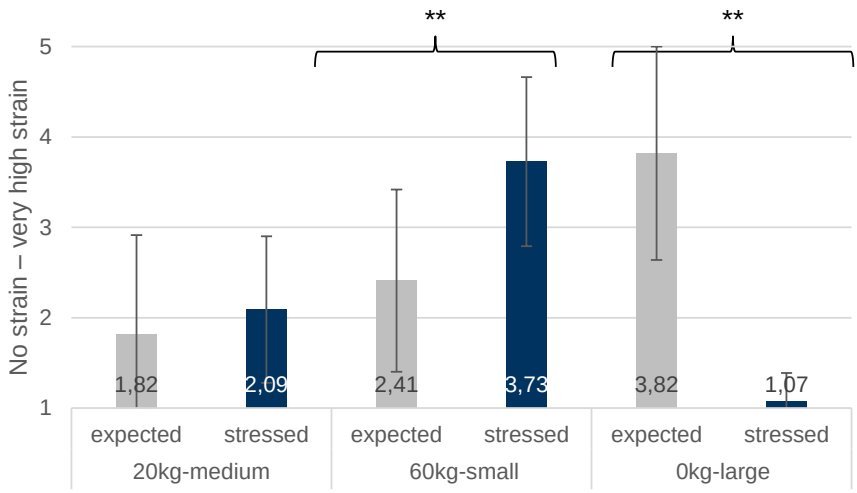


Figure 7. Expected vs. experienced strain after pushing / pulling the trolley through the parkour, for each condition

Velocity, acceleration, and jolt

The second part of the analysis is concerned with criteria to evaluate the performance of pushing / pulling tasks. In the course of this study velocity, acceleration, and jolt are considered. The median is used as a measure because of the initial mentioned insensitivity to outliers. Table 3 outlines the results arithmetically averaged over all 22 participants.

Table 3. Arithmetically averaged Mean, Standard Deviation, Median, and Maximum velocity, acceleration, and jolt for the lead-marker on the trolley between the handles

	M	*SD*	*MED*	*MAX*
velocity (m/s)	0.26	0.33	0.15	1.31
acceleration (m/s²)	2.61	2.68	3.90	47.36
jolt (m/s³)	202.82	350.54	498.62	6360.24

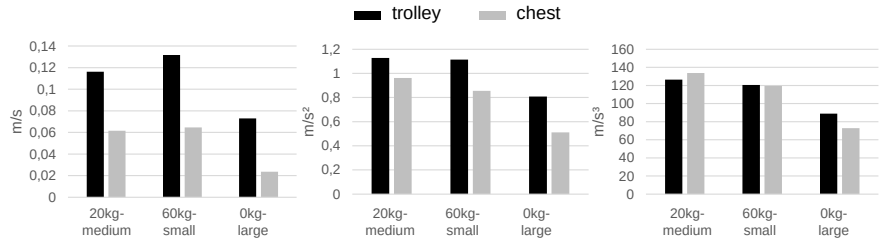


Figure 8. Arithmetically averaged median of velocity $v\left[\frac{m}{s}\right]$, acceleration $a\left[\frac{m}{s^2}\right]$, and jolt $j\left[\frac{m}{s^3}\right]$ for the trolley and chest marker for the three conditions

Figure 8 illustrates the arithmetically averaged median of velocity, acceleration, and jolt for the three states 20kg-medium, 60kg-small, and 0kg-large. Table 4 summarizes the significant influence of the weight-size-mismatch on the mentioned performance parameters. Post-hoc Bonferroni comparisons indicated that velocity, acceleration, and jolt for the conditions 20kg-medium and 60kg-small were significantly higher from 0kg-large, $p < .05$ (one exception: acceleration 20kg-medium, $p = 0.78$). The two conditions among themselves were not significantly different, $p > .05$.

Table 4. Significant influence of weight-size-mismatch on v, a, and j of the lead-marker on the trolley between the handles

	$\chi^2(2)$	ε	*F*	*df*	*p*
velocity	9.279	.694	4.609	1.386,23.610	.031
acceleration	9.364	.693	4.041	1.386,23.561	.044
jolt	8.136	.715	4.516	1.430,24.310	.032

Interpretation

The experimental design to get information about the weight-size-mismatch was implemented such that the first condition 20kg-medium was the baseline for every participant. In this way it was possible that everyone was primed to one common condition. With this in mind the estimated strain was given just on the visual impression of the object size. It is not very surprising that larger objects convey a higher estimated strain. In further investigations it will be tested how much one object size influences the operator when there are alternated weights laden.

The results of the second part suggest that higher values for v, a, and j could be indicators for better push / pull performance respectively efficiency. The very low velocity values are explainable because of the relatively short straight part of the trail. Psychophysic methods will be used to determine Detection Thresholds (DT) and Just Noticeable Differences (JND, Baird & Noma, 1978; Gescheider, 2013) for pushing / pulling tasks in further investigations.

Conclusion

In short, the study shows that the humans' expectation about feedback is highly influenced by the size of the object they have to handle. In addition to that they need enough feedback (virtual weight higher than 30N) to perform more efficiently. Humans accelerate faster (jolt), higher (a), and get to higher velocities (v) when there is needed a certain amount of force. If this requirement is fulfilled humans tend to accelerate in a comparable way.

Acknowledgments

The authors would like to acknowledge the German Federal Ministry of Education and Research for funding the project KobotAERGO (V4ARB061), in which this study was realised. We appreciate the opportunity to have carried out this study.

References

Akella, P., Peshkin, M., Colgate, E., Wannasuphoprasit, W., Nagesh, N., & Wells, J. (1999). Cobots for the automobile assembly line. *International Conference on Robotics 1999* (pp. 728–733). doi:10.1109/ROBOT.1999.770061

Baird J.C., & Nom, E. (1978). Fundamentals of scaling and psychophysics. New York: Wiley.

Bortot, D., Ding, H., Günzkofer, F., Stengel, D., Bengler, K., Schiller et al. (2010). Effizienzsteigerung durch die Bewegungsanalyse und -modellierung der Mensch-Roboter-Kooperationen. *Zeitschrift für Arbeitswissenschaft 2*, 65-75. *Retrieved from:* http://www.zfa-online.de/informationen/leser/volltexte/2010/2010_02_volltexte/Beitrag_1_ZfA_2_2010.pdf; 23.09.2014

Colgate, J. Peshkin, M. & Klostermeyer, S. (2003). Intelligent assist devices in industrial applications: a review (pp. 2516–2521). doi:10.1109/IROS.2003.1249248

Da Silveira, G., Borenstein, D., & Fogliatto, F. S. (2001). Mass customization: Literature review and research directions. *International journal of production economics*, *72*, 1-13.

DIN EN ISO, 10218-2 (2012). *Robots and robotic devices – Safety requirements for industrial robots – Part 2: Robot systems and integration*, Berlin: Beuth Verlag GmbH.

Fogliatto, F.S., da Silveira, G.J., & Borenstein, D. (2012). The mass customization decade: An updated review of the literature. *International Journal of Production Economics*, *138*, 14-25.

Gescheider G.A. (2013) *Psychophysics: the fundamentals*. Psychology Press.

Groten, R.K. (2011). *Haptic human-robot collaboration: How to learn from human dyads* (doctoral dissertation, Technische Universität München).

Kistler (2014). *Hand Force Measuring System for Ergonomics, Biomechanics and Occupational Health & Safety* (Type 9809A); *Retrieved from:* http://kistler.com/de/en/product/PSEFO/9809A; 05.09.2014

Peshkin, M., & Colgate, J.E. (1999). Cobots. *Industrial Robot: An International Journal*, *26*, 335-341.

Robotic Industries Association. (2002). *T15. 1 Draft Standard for Trial Use for Intelligent Assist Devices-Personnel Safety Requirements. Retrieved from:* http://peshkin.mech.northwestern.edu/publications/2002_T15.1_DraftStandard ForTrialUse_IntelligentAssistDevicesPersonnelSafetyRequirements.pdf; 23.09.2014

Schlick, C. M. (2009). Industrial engineering and ergonomics: Visions, concepts, methods and tools: Festschrift in honor of professor Holger Luczak. Berlin; Heidelberg: Springer-Verlag.

Seidl, A., Trieb, R., & Wirsching, H.J. (2008). SizeGERMANY – die neue Deutsche Reihenmessung–Konzeption, Durchführung und erste Ergebnisse. Produkt- und Produktions-Ergonomie–Aufgabe für Entwickler und Planer. *Gesellschaft für Arbeitswissenschaft* (pp. 391-394) GfA-Press, Dortmund.

Validation of a Telephone Manager for stressful driving situations

Linda Köhler[1], Klaus Bengler[2], Christian Mergl[3],
Kathrin Maier[4], & Martin Wimmer[1]
[1]AUDI AG, Ingolstadt
[2]Institute of Ergonomics, Technische Universität München
[3]Brose Fahrzeugteile GmbH,Coburg
[4]Institute of General Psychology, Katholische Universität Eichstätt-Ingolstadt
Germany

Abstract

Today we face highly complex urban driving situations including high information density, short decision times and a variety of stimuli acting. Crossing an intersection where drivers have to give way to crossing traffic has been identified as an example of one type of stressful situation. Several studies show that telephone calls while driving affect various aspects of driving performance. Additional stress for the driver is assumed. In order to pursue the aim of comfortable and safe driving with minimum stress even in complex situations, a suitable user interface solution including a Telephone Manager is introduced. A driving study was conducted with 27 participants validating a Telephone Manager suppressing incoming calls in stressful driving situations. Both the driving situations (turn left vs. go straight) and the telephone call (being answered vs. being suppressed) were tested towards against the driver's perceived mental workload, driving performance and acceptance. The results show a higher stress level for the driver in intersection situations. Furthermore, it confirmed that phone calls lead to additional stress, which can be reduced by call suppression in stressful situations. Moreover, the questionnaires confirmed that the telephone manager is highly accepted.

Introduction

Motivation

Complex urban driving situations are posing a big challenge in everyday car journeys.
The Cooperative UR:BAN Project, supported by the Federal Ministry for Economic Affairs and Energy, deals with such challenging settings. In the sub-project "Mensch im Verkehr", the main focus lies on the human being as an actor and scheduler in traffic with its requirements and needs. Challenging situations include, inter alia, temporary dynamics, a large number of static and moving objects, interaction with urban traffic and little space for manoeuvres.

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

In former research, crossing intersections can be identified as one of the most stressful urban driving situations (e.g. Praxenthaler, 2003; Köhler et al., 2013). T-junctions, in particular, where drivers have to give way to crossing traffic implying a high level of stress for the driver (Köhler et al., 2013). These results can be explained using cognitive psychology approaches concerning driver behaviour, described below.

Driving task and workload

In general, the driving task can be divided into three main subtasks: primary (driving process), secondary (reactions or activities deriving from the current traffic situation) and tertiary tasks (satisfaction of needs concerning the driver's comfort, information or communication) (Bubb, 2003). Furthermore, models with three hierarchy layers of the primary driving task – divided into navigation, guidance and control – have been postulated (Bubb, 2003; Donges, 1982). By splitting it into its components, it becomes apparent how complex the driving task is. This includes reaching the destination safely whilst adhering to the traffic rules. The driver has to carry out different behaviour patterns simultaneously. Cognitive demand increases for an experienced driver from the lowest level "control", via "guidance" up to "navigation" (Reichart & Haller, 1995). Rasmussen (1983) proposed the SRK taxonomy to distinguish between the different strengths of mental workload. It defines skill-based, rule-based and knowledge-based behaviour.
When merging the approach by Donges (1982) with the SRK taxonomy by Rasmussen (1983), the guidance (secondary level) and the control level (tertiary level) include skill- and rule-based activities. Based on practice and experience the driver can handle these activities – e.g. performing certain driving manoeuvres or staying with a lane – mostly unconsciously. Navigation (primary level), implies knowledge-based processes (Rasmussen, 1983), for instance perception of relevant route information. The model distinguishes between three categories with varying degrees of cognitive workload: control and guidance, in particular, are tasks which can be carried out with a low level of cognitive effort after having been learnt (rule-based processes) (Donges, 2012).Other subtasks of the primary, the secondary and the tertiary driving task follow skill- or knowledge-based modes of behaviour which place more strain on the driver's cognitive resources.
The overall construct, with regard to the availability or allocation of cognitive resources, is human attention. For the phenomenon, that attention is limited and information has to be selected, several explanatory approaches exist, two examples being bottleneck models of attention (Broadbent, 1958) and capacity models of attention (Kahneman, 1973). As De Waard (1996, p.12) proposed, on the one hand there are "concepts of a limited processing capacity" and on the other hand there are "resources calculated as the amount of processing facilities". Furthermore, the approach used to describe output losses is marginal. However, it is crucial to say that mistakes are made if too many tasks have to be fulfilled simultaneously.
In relation to the driving task, De Waard (1996, p.24) postulated an adequate model considering the driver's workload, performance and demand. The optimum is described as being a low cognitive workload that obtains a maximum result (optimal performance). By increasing demand, a higher, task-related effort will be necessary to keep the level of performance. If the demand exceeds the capacity limit the result

is mental overload. Because of the sharp rise in workload, there is a rapid decline of performance as a consequence.
In this context mental workload can be defined as "the result of reaction to demand; it is the proportion of the capacity that is allocated for task" (De Waard, 1996, p.17).
When developing advanced driver assistance systems (ADAS) and information systems, it is essential to consider the mental workload of the driver. Not least because this is encumbered by a large number of vehicle systems, followed by an even larger number of status messages. All of them are being presented to the driver in almost any situation at almost any time. So, the aim should be a minimization of workload caused by the tertiary driving task. This means that situational workload management has been developed.

Workload manager

There are many different approaches for reducing the driver's mental workload. For instance, Muigg produced an implicit workload management system. He focuses on the avoidance of non-essential driver distraction caused by messages inside the car that are inappropriate for the situation (Muigg, 2009). Another example is the information manager by Seitz (2013), which has been developed for utility vehicles. Seitz's information management system estimates the driver's current workload based on the given driving situation and the environmental conditions. Most approaches are generated, needing plenty of different pieces of information about the driver, traffic and car. In consequence, it is the aim to develop an easy to handle, easy to implement (in the car), transparent and consistent workload management system. The Information Manager by Köhler et al. (2013) describes in detail why incoming information (such as low fuel signals or windscreen washer signals) should be suppressed in stressful driving situations. Several studies show that making telephone calls while driving affect various aspects of driving performance. The driver is placed under additional stress (Tractinsky et al., 2013; Rosenbloom, 2006; Shinar et al., 2004).
An important question when considering the environment is: Will a Telephone Manager that suppresses telephone calls whilst the driver is managing stressful situations work just as well? The hypothesis is that the Telephone Manager can reduce the driver's workload while crossing an intersection and will be accepted.

Driving study

A driving study has been conducted focusing on the following questions: Can increased workload, caused by incoming calls, be proven whilst driver is managing urban scenarios? Will the suppression of incoming calls in stressful driving situations lower the level of mental workload? Will a Telephone Manager that suppresses incoming calls in stressful driving situations be accepted by the driver? In addition the validation of the intersection scenario as an example for stressful driving situations is part of the study.
Therefore, the central hypotheses are as follows: 1) A crossing situation is more stressful than going straight on. 2) A telephone call whilst driving is more stressful than no call. 3) Transferring a telephone call whilst driving in comparison to suppressing the call increases mental workload. 4) The Telephone Manager will be

accepted. Therefore, two different driving situations (crossing a T-junction by turning left vs. going straight) and three different telephone conditions (no incoming telephone call vs. call being answered vs. call being suppressed) were analysed. To standardise the contents of the telephone calls, arithmetic problems had to be solved (see also Shinar et al., 2004).
There are several methods used for measuring workload – self-report, performance and physiological measures (De Waard, 1996). In this study, performances of driving task (average speed) and telephone task (including mean time to respond to the call) (McKnight & McKnight, 1993; Shinar et al., 2004; Tractinsky et al., 2013), as well as subjective values (NASA TLX) (Hart & Staveland, 1988) were used as indicators. Personal attitudes towards the Telephone Manager were tested with the Van der Laan Acceptance Scale (Van Der Laan et al., 1997) – an instrument containing the two dimensions *usefulness* and *satisfaction*.

Materials and methods

Participants

A total of twenty seven volunteers took part in this study, being recruited through a mailing list. The sample consisted of eleven female (41%) and sixteen male (59%) participants with an average age of 35.93 years (SD_{age} = 12.7) ranged from 20 to 58 years. All of them were native German speakers in possession of a valid driving licence for at least three years (M = 17.4). 78% of the participants cover a driving distance of at least 10,000 km per year. Seventeen participants (63%) are physically able to connect their mobile phone with their private car, while 63.2% of them use this functionality at least occasionally ("occasionally" = 15.8%, "often" = 5.3%, "always" = 42.1%). Because of technical problems, two participants had to be excluded.

Apparatus

An Audi A6 Saloon with an integrated Driver Information System with 7″ colour display and a Multi Media Interface (control panel operating a separate MMI display) was used as a test vehicle. The Audi A6 had an automatic transmission. A telephone was connected to the vehicle via mobile telephone preparation with a Bluetooth interface, meaning that hands-free calls were possible using the microphone.
The whole study was conducted at the testing ground of the Universität der Bundeswehr in Munich, Neubiberg. At the testing ground urban driving scenarios were created.
To record data, both situations – crossing a T-junction whilst giving way to crossing traffic and going straight on – were tagged by trigger points which were detected by the A6 using DGPS. Both situations covered a route of 110m and were subdivided into six successive phases, as seen in Köhler et al. (2013). An Audi Q7, driven by a professional examiner, constituted the (critical) crossing traffic.

Procedure

At the start, each participant received a short briefing, including being asked to answer incoming calls while driving. The test subjects had to solve arithmetic problems,

communicated by the speaker on the telephone. For every correct calculation they would receive a bonus of 50 cents. The briefing was followed by a few manoeuvres to become familiar with the test vehicle. Whilst they got to know the Audi A6, participants received two incoming test calls – one whilst stationary and one whilst driving.

The test drive was made up of five laps of the course with each lap including one of the five test scenarios. Participants were instructed to keep a speed limit of 30 km/h, follow the traffic laws and, if they wished, to answer incoming telephone calls. The participants had to go through five scenarios (see settings in Figure 1): 1) Crossing a T-junction by turning left a) without a telephone call; b) with an incoming call (followed by an arithmetic problem); c) with a message (via Driver Information System) about a suppressed call after passing a trigger point 5 metres behind the junction. 2) going straight on for 110 metres a) without a telephone call; b) with an incoming call (followed by an arithmetic problem).

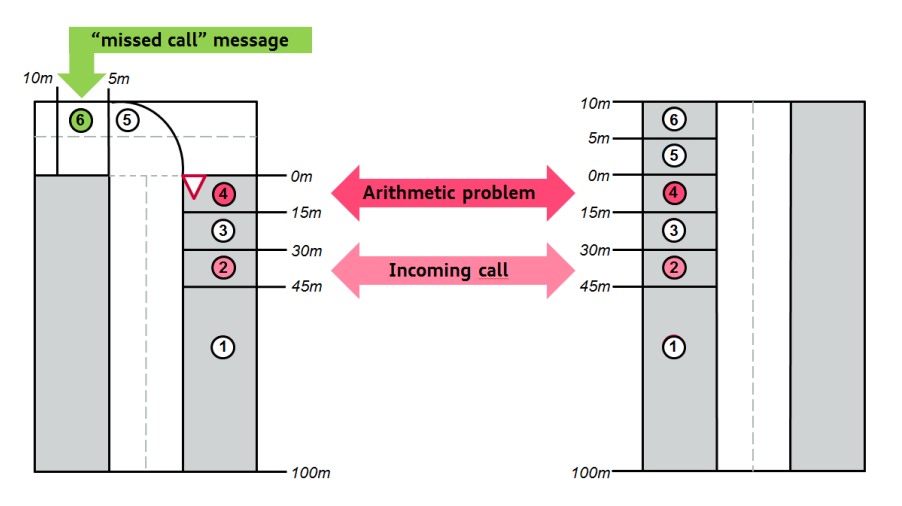


Figure 1. The two scenarios (left: turn left; right: go straight) divided into six phases with the following trigger points: incoming call (light red), arithmetic problem on the phone (red) and message about a suppressed call (green).

While crossing the intersection, the Q7 was the crossing traffic. All situations were permuted for each participant. The participant had to fill in the NASA TLX for measuring the perceived driver's mental workload after every scenario. Furthermore, in scenarios with incoming calls the examiner logged the time the participants took to answer the call and time taken to solve the arithmetic problem. At the end, the functionality of the Telephone Manager was explained to the participants. The Van Der Laan Acceptance Scale had to be completed, followed by personal information. In total, one test took about one hour and fifteen minutes per participant.

Analyses

A significance level of α=5% was assumed for testing the hypotheses. In order to

allow inferential statistics, all scales of measurement were metric. NASA TLX was adopted as recommended by Hart & Staveland (1988) ascertaining weights for each item when calculating a total amount. Recorded driving data was analysed starting from the point of a potential call (shown in figure 1). Statistical outliers were also adjusted.

A two-way repeated measure, ANOVA, was used to investigate differences in driving scenarios (turn left, go straight) and in telephone conditions (call being delivered, no telephone call). For that purpose, the amount of the NASA TLX and the average speed were used. The same measures were used for testing differences between the three telephone conditions (no telephone call, call being delivered, call being suppressed) in a univariate ANOVA with repeated measures. To compare all three telephone conditions (no incoming telephone call, call being answered, call being suppressed), a t-test (predisposed individual comparisons) was used for testing subjective and objective data. The mean time to respond to the call and the mean time to solve the arithmetic problem were compared for the scenarios *turning left* and *going straight* using a t-test for paired samples. A t-test for paired samples was used to find the difference between the two telephone conditions (telephone call while driving, no telephone call). Finally, the acceptance of driving with the functionality of the Telephone Manager and without the functionality was compared by means of a t-test. The subscales of *usefulness* and *satisfaction* have been calculated for this.

Results

The subjective evaluation concerning drivers' mental workload shows no difference between *turn left* (M=19.15; SD=14.07) and *go straight* (M=15.68; SD=13.56). Even though there was no significant main effect for the subjective amount of the NASA TLX, $F(1,25) = 3.65$, $p = .07$, $\eta^2_p = .13$, ns., a tendency emerged, approved by the p-value and the effect size. This trend has been confirmed by the mean time to respond to the incoming call – while crossing the T-junction (M=2.4s; SD=0.82s) participants took significantly longer to respond compared with going straight (M=2.16s; SD=0.78s), $t(21) = -1.73$, $p < .05$ (Figure 2). However, the mean time to solve the problem on the phone did not differ significantly, $t(21) = 0.97$, $p > .05$, ns. For calculating participants needed as much time by turning left (M=3.19s; SD=4s) as by going straight (M=4.25s; SD=4.51s).

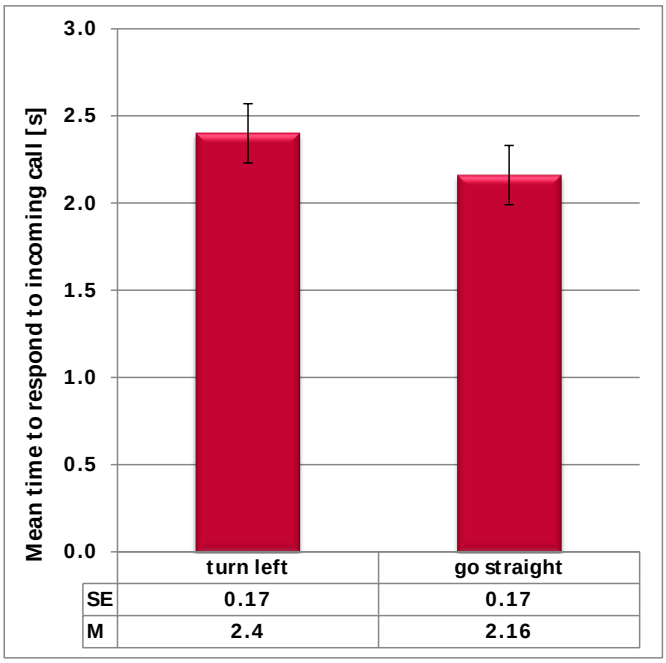


Figure 2. The two scenarios (turn left and go straight) compared by the mean time it took the participants to answer an incoming telephone call. The difference is statistically significant.

By comparing conditions with and without an incoming call, a significant effect can be shown using NASA TLX, $F(1,25) = 23.69$, $p < .001$. Without a phone call participants stated lower mental workload (M=9.53; SD=10.25) in comparison to answering an incoming call while driving (M=25.3; SD=19.22). The average speed did not depend on the telephone condition, $F(1,20) = 0.97$, $p > .05$, ns. Nevertheless, by considering individual comparisons for crossing the intersection, according to the hypothesis, deviations in the average speed with (M=19.25km/h; SD=3.2km/h) and without phone call (M=20.87km/h; SD=2.31km/h) were significant, $t(23) = 5.02$, $p < .001$. For driving straight on it did not show any deviation, $t(21) = -0.41$, $p > .05$, ns.

Comparing the scenario *intersection*, the three different telephone call conditions differed significantly, $F(2,50) = 14.55$, $p < .001$ (Figure 3). *Answered call* shows the highest level of mental workload (M=27.04; SD=20.51), by contrast to *call being suppressed* (M=13.49; SD=12.98), $t(25) = 3.74$, $p < .001$, and *no incoming call* (M=11.27; SD=11.9) which are almost equal, $t(25) = -1.09$, $p > .05$, ns.

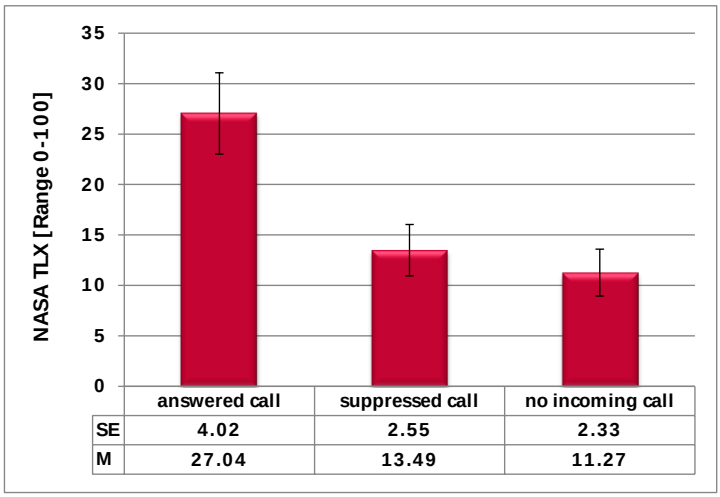


Figure 3. The three different telephone conditions (no incoming telephone call vs. call being answered vs. call being suppressed) at the scenario "turn left" compared by their level of mental workload (NASA TLX).

Objective data gave proof of this effect, as well. The average speed was significantly concerning the factor "telephone call", $F(2,46) = 14.19$, $p < .001$. During an *incoming call* (M=19.38km/h; SD=3.21km/h) in comparison to the scenario with a *suppressed call* (M=20.86km/h; SD=2.26km/h), participants drove significantly slower, $t(24) = -3.51$, $p = .001$. There was no difference measured between *suppressed call* and *no call* (M=20.87km/h; SD=2.31km/h), $t(23) = 0.06$, $p > .05$, ns.
The Van Der Laan Acceptance Scale is able to assess system acceptance in two dimensions – a Usefulness Scale and a Satisfying Scale.
Comparing the Usefulness Score, a significant difference between a car with the functionality of a Telephone Manager (M=-0.84; SD=1.0) and without the functionality (M=-0.2; SD=0.92) has been shown, $t(26) = -2.14$, $p < .05$ (Figure 4).

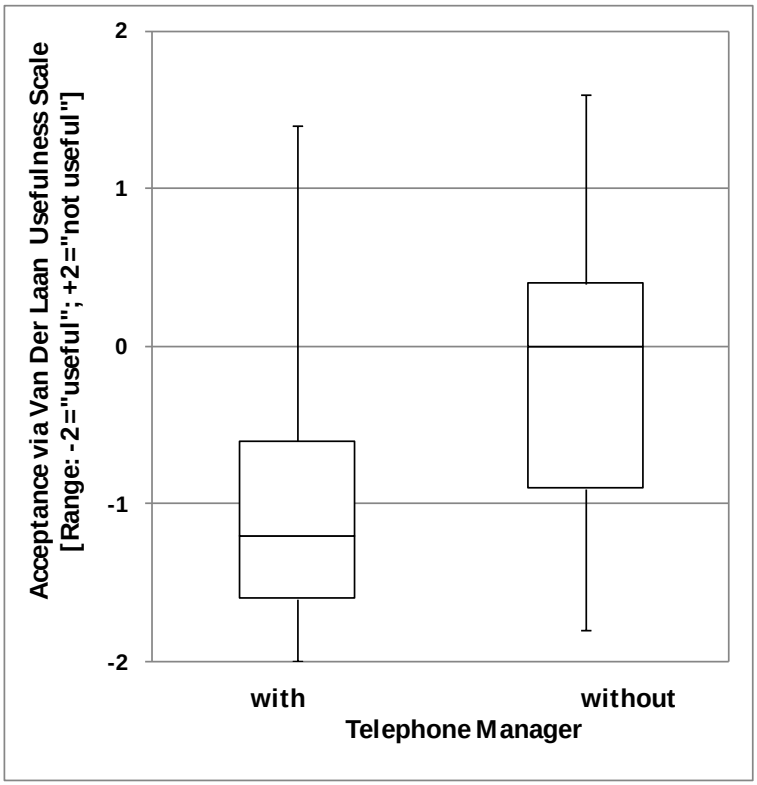


Figure 4. Acceptance of suppressing calls in stressful driving situations (Telephone Manager) on the basis of the Usefulness Scale as part of the Van Der Laan Scale (Van Der Laan et al., 1997).

The comparison of the Satisfying Score showed statistically significant differences, $t(26) = -3.16$, $p < .01$. The Telephone Manager (M=-0.89; SD=0.98) is evaluated as being more satisfying than a car without the functionality (M=0.13; SD=1.13).

Discussion

The study aimed to confirm the Telephone Manager as a function that decreases workload in stressful driving situations. The Manager was implemented by suppressing incoming phone calls while the driver had to handle a left turn at a T-junction and give way to crossing traffic. In detail the functionality is suppressing incoming calls in phases of high driver's mental workload (compare Figure 1: phases of high driver's mental workload are phase 2, 3, 4 and 5).
First of all, crossing the intersection had to be identified as a stressful driving situation. The first hypothesis expects a higher workload for the scenario *turn left* in comparison to the scenario *go straight*. Subjective data (NASA TLX) showed a small tendency but no statistical significance. An identical effect can be shown with the mean time of solving the arithmetic problem on the phone. Only the mean time to respond to an incoming call confirmed the hypothesis. Referring to Rasmussen's classification

(1983) *going straight* relies on skill-based processes (guidance and control); as opposed to *crossing the intersection*, which requires rule-based processes and therefore demands cognitive control. More *time taken to respond to the call* indicates that more attention is needed to manage the primary driving task (Rasmussen, 1983). Longer processing times are a result of the apportionment of mental resources split through driving task and secondary task (Kahnemann, 1973). During the easier scenario (going straight) the telephone ringing was captured earlier. An explanation is the availability of more capacities for the secondary task (resource models) or the lower charged processing channel (1-channel-model) (De Waard, 1996). The environmental conditions at the testing ground in Neubiberg were causing only a low level of mental workload for the driver in general. There were no pedestrians, no cyclists and one Audi Q7 forming the crossing traffic. Transferred to urban traffic situations, differences in workload will rise up as shown by Köhler (2013). Besides, NASA TLX scores showed high values of standard deviation. This can be explained by the small number of participants.

The second hypothesis relates to mental workload caused by telephone calls while driving a car. On the subjective level it can be proven that telephone calls increase drivers' mental workload in both scenarios. On the objective level the impact merely appears to be at the intersection. In this scenario, participants reduce speed when making a telephone call. Compared to going straight, where the average speed does not depend on incoming calls. This phenomenon can be interpreted by reference to the keynote by De Waard (1996). The fact that performance declines in the intersection scenario but not in the *going straight* scenario – even if NASA TLX shows a high level for both of them – can be explained by the region model (Figure 5; De Waard, 1996, p.24).

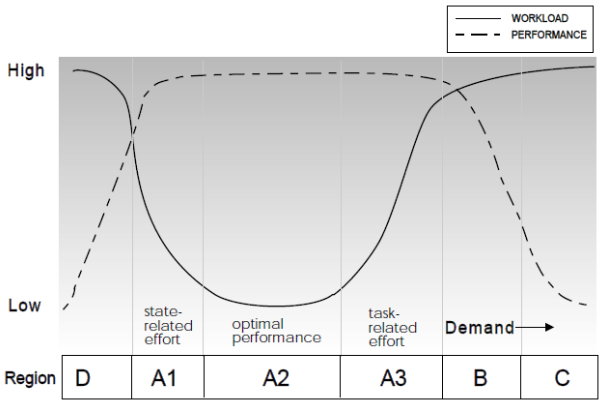


Figure 5. Region model by De Waard (1996, p. 24) depicting the relation between demand, workload and performance in 6 regions.

As shown in Figure 5 and referring to theoretical assumptions, region A3 can be characterised as follows: "[...] performance measures still do not show a decline, but the operator is only able to maintain the level of performance by increasing effort." (De Waard, 1996, p. 23). This is consistent with the scenario *going straight* and answering an *incoming telephone call* – even if driving performance (average speed)

doesn't show an impact of the phone call, subjectively the mental workload increases (NASA TLX). Compared to the second scenario (making a phone call whilst crossing the intersection), driving performance is affected, as shown in region B (De Waard, 1996). In this context, performance deficits can be explained based on limited resources. Crossing an intersection was identified as a rule-based action, needing more processing capacity than going straight. Because resources have to be shared for the incoming call, driving performance deficits arise (Rosenbloom, 2006; Shinar, 2004).

For confirming the Telephone Manager by disclosing its benefits, a third hypothesis was defined to identify a decrease in mental workload caused by the function. The Telephone Manager suppresses incoming phone calls in stressful driving scenarios. In the study "crossing the intersection" was used as an example for such situations. The results confirm a decrease in the driver's mental workload when calls were suppressed compared to answered calls. The subjective evaluation (NASA TLX) as well as objective data (average speed) identified a significantly higher level of mental workload when calls are answered in the stressful driving scenario "intersection" (turn left). Suppressed calls show a low level of workload as well as the condition "no call". Because of the suppression of the call, additional workload can be prevented. By consequence, all processing capacities will be available for managing the driving scenario.

A fourth hypothesis was put forward to confirm whether the Telephone Manager will be accepted by the driver. The validated Acceptance Scale by Van Der Laan yields a significant impact in the Usefulness Scale and the Satisfying Scale. Participants prefer the new functionality for stressful driving situations. The Telephone Manager is accepted.

Conclusion

In brief, the study shows that telephone calls while driving cause a higher mental workload. Also, the Telephone Manager – suppressing incoming calls in stressful driving situations – decrease the level of drivers' workload level significantly. Even though crossing an intersection couldn't be identified as such a stressful scenario, workload can be lowered here as well. Besides, the developed concept will be accepted by the driver. In this context, it is important to note, that the stressful driving scenario usually does not take longer than thirty seconds. Hence, there are only a few occasions where an incoming call will be suppressed entirely. A solution could be to only suppress the initial ringing.

In summary, this study shows the usefulness of the Telephone Manager and encourages its introduction for stressful driving situations. As this functionality just bases on predictive road data, its implementation will be less complicated compared to other Workload manager approaches, which require a more complex technical infrastructure like interior sensors, on-board network or bus data.

References

Broadbent, D.E. (1958). *Perception and communication*. London: Pergamon Press.

Bubb, H. (2003). Driver assistance: firstly a contribution to primary safety or rather to comfort? In VDI –Gesellschaft Fahrzeug- und Verkehrstechnik (Ed.), *7.*

Tagung – Der Fahrer im 21. Jahrhundert Fahrer, Anforderungen, Anwednungen, Aspekte für Mensch-Maschine-Systeme (VDI Berichte 1768) (pp. 25-46). Düsseldorf: VDI Verlag GmbH.

De Waard, D. (1996). *The measurement of drivers' mental workload.* PhD thesis, University of Groningen. Haren, The Netherlands: University of Groningen, Traffic Research Centre.

Donges, E.(1982). Aspekte der Aktiven Sicherheit bei der Führung von Personenkraftwagen. *Automobil-Industrie, 27,* 183-190.

Donges, E. (2012). Fahrerverhaltensmodelle. In H. Winner, S. Hakuli, and G. Wolf (Eds.), *Handbuch Fahrerassistenzsysteme. Grundlagen, Komponenten und Systeme für aktive Sicherheit und Komfort* (pp. 15-23). Wiesbaden: Vieweg + Teubner.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.) *Human mental workload* (pp. 139-183). Amsterdam: North-Holland.

Kahneman, D. (1973). *Attention and Effort.* Englewood Cliffs, USA: Prentice Hall.

Köhler, L., Mergl, C., Blaese, D., & Bengler, K. (2013). Fahrerbeanspruchung im urbanen Raum: Erhebung der subjektiven Beanspruchung des Fahrers bei einer Kreuzungsüberquerung. In VDI –Gesellschaft Fahrzeug- und Verkehrstechnik (Ed.), *7. Tagung – Der Fahrer im 21. Jahrhundert Fahrer, Fahrerunterstützung und Bedienbarkeit (VDI Berichte 2205)* (pp. 237-250). Düsseldorf: VDI Verlag GmbH.

McKnight, A.J., & McKnight, A.S. (1993). The effect of cellular phone use upon driver attention. *Accident Analysis and Prevention, 25,* 259-265.

Muigg, A. (2009). *Implizites Workloadmanagement – Konzept einer zeitlich-situativen Informationsfilterung im Automobil.* PhD thesis, Technische Universität München. Göttingen, Germany: Technische Universität München, Faculty of Electrotechnical Engineering and Information Technology.

Praxenthaler, M. (2003). *Experimentelle Untersuchung zur Ablenkungswirkung von Sekundäraufgaben während zeitkritischer Fahrsituationen*. PhD thesis, Universität Regensburg. Regensburg, Germany: Universität Regensburg, Institute of Experimental Psychology.

Rasmussen, J. (1983). Skills, Rules and Knowledge: Signals, Signs and Symbols, and other Distinctions in Human Performance Models. *IEEE Transactions on Systems, Man and Cybernetics, 13,* 257-266.

Reichart, G., & Haller, R. (1995). Mehr aktive Sicherheit durch neue Systeme für Fahrzeug und Straßenverkehr. In W. Fastenmeier (Ed.), *Mensch – Fahrzeug – Umwelt: Vol. 33. Autofahrer und Verkehrssituation. Neue Wege zur Bewertung von Sicherheit und Zuverlässigkeit moderner Straßenverkehrssysteme* (pp. 199-216). Köln: TÜV Rheinland.

Rosenbloom, T. (2006). Driving performance while using cell phones: An observational study. *Journal of Safety Research, 37,* 207-212.

Seitz, M. (2013*). Information management in commercial vehicles.* PhD thesis, Technische Universität München. Germany, Technische Universität München, Faculty of Mechanical Engineering.

Shinar, D., Tractinsky, N., & Compton, R. (2004). Effects of practice, age, and task demands, on interference from a phone task while driving. *Accident Analysis and Prevention, 37,* 315-326.

Tractinsky, N., Ram, E.S., & Shinar, D. (2013). To call or not to call – That is the question (while driving). *Accident Analysis and Prevention, 56,* 59-70.

Van Der Laan, J.D., Heino, A., & De Waard, D. (1997). A simple procedure for the assessment of acceptance of advanced transport telematics. *Transportation Research Part C: Emerging Technologies, 5*, 1-10.

Anger and bother experience when driving with a traffic light assistant: A multi-driver simulator study

Lena Rittger, Dominik Muehlbacher, Christian Maag, & Andrea Kiesel
Adam Opel AG, IZVW, WIVW GmbH, Universität Würzburg
Germany

Abstract

Drivers evaluated their interaction with others when driving with a traffic light assistant. In a multi-driver simulator setting, four drivers drove at the same time in the same virtual environment. Two drivers were equipped with a traffic light assistant that recommended driving speed and required action, e.g. 'brake to 30 km/h'. Additionally, the position of the drivers in the column, the distance to the traffic light at which the recommendations started, and the instruction whether drivers 'can' or 'must' follow the recommendations were varied. Drivers with assistant pulled a lever at the steering wheel to indicate their feeling of bothering others. They did so most often when the assistant recommended coasting at far distances to the traffic light, especially when driving in the front positions of the column and when the instruction was that they 'can' follow the recommendations. Drivers without assistant pulled the lever at the steering wheel to indicate their anger about others. They did so only when they were following drivers with traffic light assistant. The results will help to parameterise the traffic light assistant regarding when and how to recommend.

Introduction

Modern traffic light assistance systems enable communication between infrastructure and vehicles. For example, approaching vehicles receive information about current and next state of a traffic light and about phase durations. Based on this information, the assistance system calculates driving recommendations for passing the intersection at a green light. In case of unavoidable stops, the system recommends an efficient stop at red. The main targets of the assistant are reducing emissions, increasing traffic flow and improving driver comfort (Thoma et al., 2007).

To develop driver assistance systems two goals are crucial. First, the efficiency of the system should be maximized. The degree of impact the system has on consumption, emissions and traffic flow is determined by various parameters. For example, previous research using traffic simulation tools showed that increasing the start distance for the activation of a traffic light assistant from 200 to 400 metres in front of the traffic light has beneficial effects on emissions (Tielert et al., 2010).

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Second, conditions should be created in which system behaviour is consistent with drivers' desired behaviour (Tango, & Montanari, 2006). The assistant needs to be designed in a way that maximizes comfort, acceptance and willingness to use the system. As an evaluation criterion, the emotional climate has been evolved (Maag, 2013). It can be hypothesised that emotional reactions of drivers and the expectations on emotional reactions of surrounding drivers influence the acceptance for a driver assistance system.

The current study is based on the assumption that even with a fast introduction of a traffic light assistant to the market, the penetration rates will be mixed for several years. Hence, road users driving with assistant system will interact with road users who are not equipped with traffic light assistance. This leads to a discrepancy of knowledge that drivers have of the upcoming right of way rules at the intersection: While drivers without assistant evaluate the required driving behaviour (accelerating for proceeding or decelerating to initiate a stop) only based on the current visible traffic light state, drivers with assistant initiate driving behaviour based on driving recommendations that consider time and state of the next traffic light phase. Hence, different drivers approaching the same intersection come to different conclusions on appropriate driving behaviour, based on different quality of the available information.

For road users driving without assistance system, the diverging driving behaviours potentially induce negative emotional reactions. For example, a discrepancy between desired driving speed and actual driving speed comes along with the experience of anger (Stephens & Groeger, 2014). This should be avoided, because research has pointed out that drivers experiencing anger are likely to engage in dangerous driving behaviours (Deffenbacher et al., 1994, Guéguen et al., 2014, Shinar, 1998, Stephens & Groeger, 2014). At the same time, for road users driving with assistance systems the deviation from normal driving behaviour might lead to expectations about bothering other road users. As a result, compliance to the system recommendations could be decreased. Hereby, instructions whether a driver should (must) or can follow system recommendations might influence the extent of the feeling of bothering others.

In summary, the main research questions of the present study were: Under which situational circumstances and system states do participants driving with a traffic light assistant feel that they are bothering other road users? Under which situational circumstances and system states do participants driving without traffic light assistant express that they feel angered by other road users?

Methods

Participants

44 participants (18 female) took part in the study. Due to technical problems in one session, data of 40 participants were analysed. The mean age was 38.6 years (sd = 15.8). All participants were trained for driving in the multi driver simulator. No driver had experience with a traffic light assistant.

Apparatus

The study took place in the static multi driver simulator at WIVW GmbH (Wuerzburg Institute for Traffic Sciences). At the four driving stations of the multi driver simulator four drivers drove at the same time in the same virtual road environment. Each driving station consisted of three 22" LCD displays with a resolution of 1680x1050 pixels, offering a 150° horizontal field of vision. The left display showed the field of vision experienced in the left window, including the left side mirror. The windscreen view is displayed in the middle and right display, including the centre mirror and the left side mirror, as well as the instrument cluster with speedometer. The left, front and right mirrors were depicted with a size of 11x6 cm. For the HMI of the traffic light assistant, there was an additional 10" LCD Display with 800x400 pixels positioned next to the steering wheel. As mock-ups steering wheels enhanced by force feedback and ordinary pedal systems were used. The steering wheels had two levers, one at the left and one at the right side. The simulator was run by the SILAB software.

Traffic light assistant

The algorithm of the traffic light assistant considered the current and next traffic light phase and participants' driving speed and distance to the traffic light. Based on that, driving recommendations were calculated, which contained a combination of action and speed suggestions. Action recommendations were either coast, brake or drive. Speed recommendations were either 0, 20, 30 km/h. The thresholds for the activation of the recommendations was 5 km/h, e.g. a recommendation to drive 20 km/h was presented as long as participants drove between 15 and 25 km/h. The recommendations were presented in text form with distinctive colours (Figure 1).

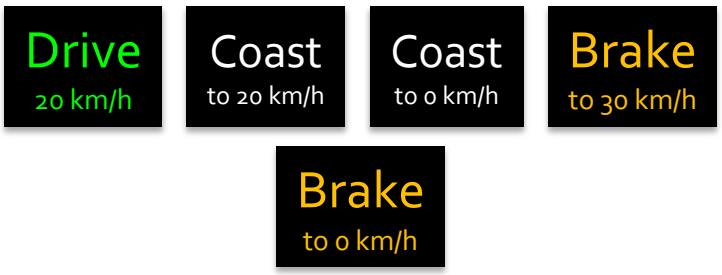


Figure 1. Driving recommendations as shown in the HMI. The drive recommendation was depicted in green, coast recommendations in white and brake recommendation in amber.

Study design

The study had a mixed between-within subjects design. Participants always drove in columns of four drivers. In each column, two of the four drivers received recommendations from the traffic light assistant system, whereas the other two drove without system. Drivers without system did not know about the existence of the traffic light assistant. Half of the drivers with system were instructed to always stick to the recommendations ('must' condition), whereas the other half of the drivers with system were instructed that they could stick to the recommendations

whenever they wanted (‘can’ condition). In the column of four drivers, each driver had four possible positions. Four different orders were realised in the experiment. The four orders ensured that each participant drove at each of the four positions for an equal number of times, that drivers with assistant system only followed drivers without assistant system and that the combinations of lead and following vehicle varied (Figure 2). Recommendations of the traffic light assistance either started at 200 m or at 400 m in front of the intersection. To investigate the influence of system activation on the dependent variable, the traffic light approach was separated into the distance sections 0 – 200 and 200 – 400 metres in front of the intersection.

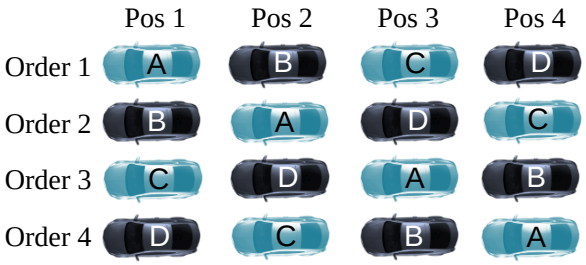


Figure 2. The four realised orders with drivers A-D in different positions of the column. Drivers A and C received recommendations from the traffic light assistant. Drivers B and D did not drive with traffic light assistant.

Drivers receiving recommendations from the traffic light assistant were instructed to pull a lever at the steering wheel every time they felt like bothering other drivers. Drivers who did not receive recommendations from the traffic light assistant were instructed to pull the lever every time they were angered by other drivers.

Procedure

Each participant was instructed individually and drove a short practice track. During the experiment four participants drove in the same virtual environment. They crossed 16 traffic light intersections without turn, which resulted from a repetition of the eight different conditions (two start distances x four column positions). The traffic light approaches were about 600 metres long. In all traffic light approaches, drivers had to reduce speed to either cross the intersection at green without stop or to initiate an efficient stop at red. Before each traffic light approach, the order of the vehicles in the column was changed. After completing all traffic light approaches, drivers filled in a short questionnaire, which is reported in the results section.

Results

Feeling of bothering others expressed by drivers with system

An analysis of variance (ANOVA) was conducted with the between-subjects variable instruction (‘can’ vs. ‘must’) and the within-subject variables notification distance (200 vs. 400 metres), position in the column (1 vs. 2 vs. 3 vs. 4) and distance section during the approach (0-200 vs. 200-400 metres). Only data of participants driving with assistant were included. The number of traffic light

approaches with lever pull was related to the total number of approaches in the respective condition and considered as dependent variable. Results are presented in table 1.

Table 1. Summary of ANOVA results for the percentage of lever pulls to express the feeling of bothering others. Bold numbers mark significant effects.

Effect	*Df effect*	*Df error*	*F*	*p*	$\eta^2_{partial}$
Instruction (I)	**1**	**18**	**20.098**	**<.001**	**.528**
Notification distance (ND)	**1**	**18**	**36.699**	**<.001**	.671
Position (P)	**3**	**54**	**12.203**	**<.001**	**.404**
Distance section (S)	1	18	3.860	.065	.177
ND x I	1	18	3.315	.085	.156
P x I	3	54	1.713	.175	.087
S x I	1	18	.095	.762	.005
ND x P	3	54	.816	.491	.043
ND x S	**1**	**18**	**18.051**	**<.001**	**.501**
P x S	**3**	**54**	**3.195**	**.031**	**.151**
ND x P x I	3	54	.420	.739	.023
ND x S x I	**1**	**18**	**8.294**	**.009**	**.315**
P x S x I	3	54	.133	.94	.007
ND x P x S	3	54	1.825	.154	.092
I x ND x P x S	3	54	1.069	.370	.056

Drivers expressed more often the feeling of bothering others in the ‘can’ condition compared to the ‘must’ instruction. When the recommendations started 400 metres in front of the intersection, drivers more often expressed the feeling of bothering others compared to when recommendations started 200 metres in front of the intersection (figure 3).

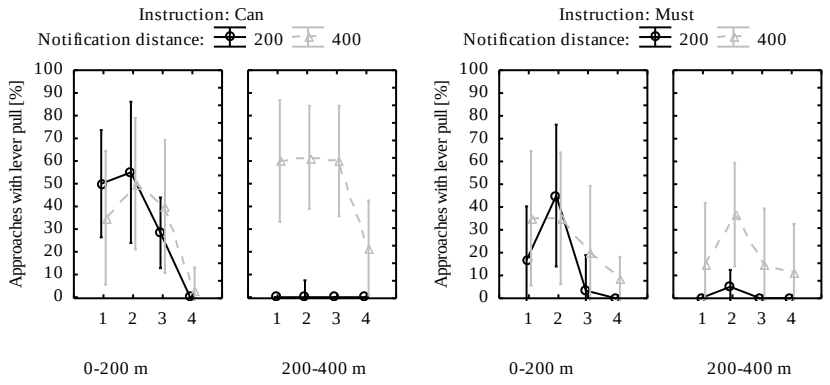


Figure 3. Percentage of traffic light approaches with lever pull expressing the feeling of bothering others by participants driving with assistance system related to the conditions position in the column, notification distance and distance section in front of the intersection. Graphs show means with 95% confidence intervals.

When drivers were in the fourth position of the column, the lever was pulled significantly less often compared to when driving in any other position of the column (all *p's* <.028). When recommendations started 200 m in front of the intersection, hardly any driver pulled the lever between 200 and 400 metres.

Anger expressed by drivers without system

An ANOVA was conducted with the between-subjects variable instruction ('can' vs. 'must') and the within-subject variables notification distance (200 vs. 400 metres), position in the column (1 vs. 2 vs. 3 vs. 4) and distance section during the approach (0-200 vs. 200-400 metres). The variables instruction and notification distance were varied for drivers with system and the impact of the variations was assessed for drivers without system. For every participant driving without assistant, the number of traffic light approaches with lever pull was related to the total number of approaches in the respective condition and considered as dependent variable. Results are presented in table 2.

Table 2. Summary of ANOVA results for the percentage of lever pulls to express anger. Bold numbers mark significant effects.

Effect	*Df effect*	*Df error*	*F*	*p*	$\eta^2_{partial}$
Instruction (I)	1	18	3.728	.069	.172
Notification distance (ND)	**1**	**18**	**15.886**	**<.001**	**.469**
Position (P)	**3**	**54**	**11.389**	**<.001**	**.388**
Distance section (S)	1	18	4.366	.051	.195
ND x I	1	18	2.179	.157	.109
P x I	3	54	2.516	.068	.123
S x I	1	18	4.366	.051	.195
ND x P	**3**	**54**	**3.928**	**.013**	**.179**
ND x S	**1**	**18**	**10.407**	**.005**	**.366**
P x S	3	54	1.672	.184	.085
ND x P x I	3	54	.432	.737	.023
ND x S x I	1	18	.15	.703	.008
P x S x I	3	54	.786	.507	.042
ND x P x S	**3**	**54**	**3.747**	**.016**	**.172**
I x ND x P x S	3	54	1.203	.318	.063

Drivers were more angered by others when the recommendations started 400 metres in front of the intersection compared to a start at 200 metres. They expressed less anger, when driving in the first position of the column compared to the second, third or fourth position of the column (all *p's* < .006). When the recommendations started 200 metres in front of the intersection, hardly any driver pulled the lever between 200-400 metres in front of the intersection (Figure 4).

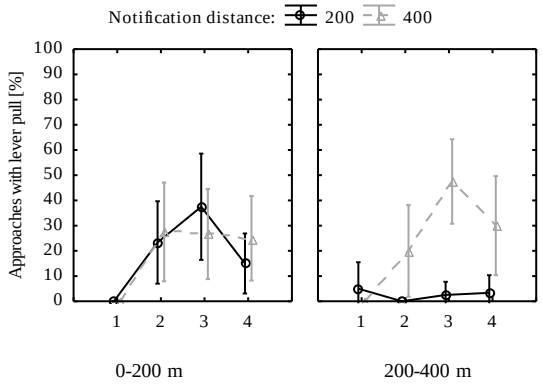


Figure 4. Percentage of traffic light approaches with lever pull to express anger by drivers without assistant related to the conditions notification distance, position in the column and distance section in front of the intersection. Graphs show means with 95% confidence intervals.

Relation of bother and anger feelings

In order to gain insight in the relation between lever pulls of drivers with and without system the number of approaches with lever pulls of drivers following each other was investigated. First, the number of approaches with lever pulls for drivers with assistant in the first, second and third position of the column was determined based on data for the overall approach distance of 400 metres. Second, from these approaches the number of approaches was identified in which the directly following driver in the second, third and fourth position also pulled the lever. By allocating both numbers it was determined in how much percent of the approaches in which a driver with system pulled the lever, the directly following driver without system expressed that he felt angered. Data are presented in table 3 for the three pairs: driver on position one followed by driver on position two, driver on position two followed by driver on position three and driver on position three followed by driver on position four. Drivers with assistant had the respective lead position, drivers without assistant had the respective following position.

Table 3. Number of approaches with lever pull of a driver with system in the first three positions of the column and percentage of approaches in which the directly following driver also pulls the lever.

Independent variable		*Number of approaches with lever pull of drivers with system []*			*Proportion of approaches with pairs pulling the lever [%]*		
Instruction	*Notification distance*	*Position 1*	*Position 2*	*Position 3*	*Pair 1/2*	*Pair 2/3*	*Pair 3/4*
'Can'	200	12	9	8	41.66	33.33	0.00
	400	12	21	11	33.33	42.85	36.36
'Must'	200	4	6	1	0.00	83.33	0.00
	400	7	13	5	57.14	76.92	80.00

Overall, in 40.413% of the cases in which drivers with assistant expressed that they bothered others the following drivers also expressed feeling angered by others.

Dependence on type of recommendation

A further analysis was conducted to investigate the number of lever pulls depending on the five different driving recommendations. The total time at which the specific recommendation was presented during all 16 traffic light approaches was determined for each participant driving with the system (figure 5, dashed line). The long durations of the 'brake to 0' recommendations were measured in cases when the assistant did not turn off in standstill when waiting at red traffic lights. Additionally, the number of episodes with at least one lever pull occurring while each of the recommendations was active was identified. For each participant, the number of episodes with lever pull was related to the total time spent with activated recommendation. The resulting ratios are presented in figure 5. A within subject ANOVA was conducted with recommendation as independent variable and the ratio as dependent variable. The ratio differed significantly for the recommendations, $F(4,76) = 11.409$, $p < .001$. $\eta^2_{partial} = .375$. Bonferroni adjusted post-hoc tests showed that the 'coast to 20 km/h' recommendation led to significantly more lever pulls compared to all other recommendations, all *p's* < .033. Additionally, the 'coast to 0 km/h' recommendation led to more lever pulls compared to the 'brake to 0 km/h' recommendation, $p = .016$.

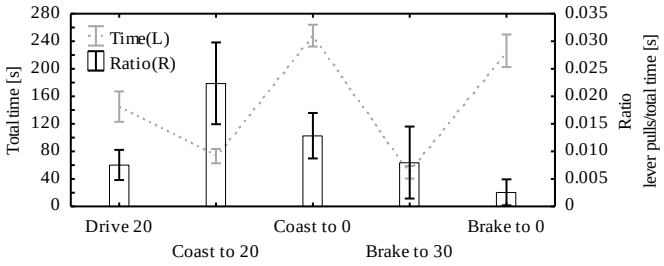


Figure 5. Total activation time (left axis) and number of lever pulls in relation to total time spent with activated recommendation of the traffic light assist (right axis) related to the five recommendations. Graph shows means with 95% confidence intervals.

Questionnaire

Drivers without assistant were asked if they still felt anger in case they knew about the assistance system other drivers are using (e.g. by a sticker at the back of the car). Drivers with assistant were asked if they still felt like bothering others when driving with the assistant, in case others would know about their system (e.g. by a sticker at the back of their own car). Figure 6 indicates participants' agreement to these statements. There was no difference between drivers with and without assistant, $p = .917$.

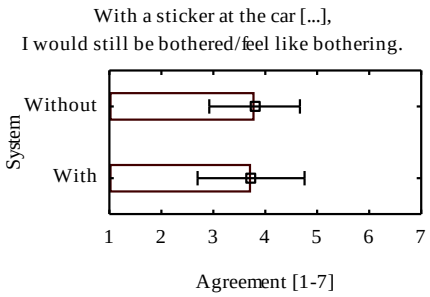


Figure 6. Drivers' agreement to the statement asking for a change of anger or bother experience in case others knew about the traffic light assistant system. 1 indicates that their bother would have been reduced. 7 indicates that their anger would have been the same. Graph shows means with 95% confidence intervals.

Discussion

The influence of traffic and system parameters on drivers' bother and anger experience when driving with a traffic light assistant was investigated. The traffic light assistant allows drivers to pass intersections at a green light or to initiate an efficient stop at red. It was expected that the assistance system triggers negative emotional reactions in relation to interactions between road users equipped with the assistant and un-equipped drivers. The multi-driver simulator allows for investigating interactions between real drivers in a controlled laboratory setting.

The results show that the traffic light assistant has the potential to induce anger in drivers without assistant and the feeling of bothering others in drivers with assistant. Drivers with system especially felt like bothering others in the front position of the column. Drivers without assistant were especially bothered when driving in the back positions of the column. The analysis revealed that drivers with assistant more often expected to bother other drivers than the directly following drivers expressed that they were angered by others. However, it is the expectation on negative reactions by others that might reduce compliance when driving with the assistant and with that might lower possible benefits of the system. Therefore, future research could investigate how the deviation between expectations on others negative reactions and the actual arising emotions could be used in order to motivate drivers to feel comfortable when using the system.

A simple solution might be to inform others about the traffic light assistant in the vehicles. Research has shown that anger in others can be larger when drivers do not see the reasons for reductions in driving speed of a lead vehicle (Stephens & Groeger, 2014). Drivers responded that the sticker at the back of the car could have some potential to reduce anger and bother. The sticker could reduce the feeling of being limited in the free choice of speed in drivers without system and emphasise that even without system one can benefit from following a lead vehicle with assistant (e.g. in avoiding a stop at red). Future research could address if the egocentric perspective that drivers have when interacting in traffic could be

improved by more information exchange and elucidation on other drivers' motives and backgrounds.

Unlike expected, the instruction that drivers 'can' stick to the recommendations led to an increased likelihood for lever pulls in drivers with assistant compared to the instruction to always stick to the recommendations ('must' condition). An explanation for that might be that drivers in the 'can' condition complied to the recommendations voluntarily, but wanted to express that they are not confident with the recommendations. Drivers in the 'must' condition had no choice and therefore contributed their cumbersome behaviour to the system.

In the 0-200 metres in front of the intersection, drivers pulled the lever equally often, independent of the start of the driving recommendations. When recommendations started 200 metres in front of the intersection, hardly any driver expressed the feeling of anger or bothering others 200-400 metres in front of the intersection. Hence, lever pulls were related to system activation. Additionally, it shows that when recommendations started at far distances to the intersection, anger or bother only slightly reduced over the course of the approach. Therefore, the 400 metres notification distance condition has a higher potential to trigger anger or bother feelings in drivers. Along with that, the coast recommendations led to the highest number of bother episodes. In the 'coast to 20 km/h' and 'coast to 0 km/h' recommendations the deviations from the maximum speed limit were largest. The reason for a lower number of bother experiences in the 'drive 20 km/h' recommendation could be that the recommendations were presented consecutively during each approach. It might be that drivers expressed their feeling of bothering others in the preceding 'coast to 20 km/h' situation and did not repeat it afterwards in the 'drive 20 km/h' condition. For the parameterisations of the traffic light assistant it is important to aim for a trade-off between maximum efficiency and maximum driver acceptance. Even though the traffic light assistant is more efficient when activated at far distances to the intersection and with initiating long coasting episodes, the benefits for comfort, emissions and efficiency of traffic flow will be reduced, when drivers feel uncomfortable in using the system.

A possible flaw of the method of lever pulls is that drivers are explicitly instructed to express their negative feelings in the interaction with others. Therefore, the setting could emphasise the negative effects of driving with the system and might overestimate drivers' anger and bother experience. For future research it is recommended to compare the current results to other measures of anger or discomfort (e.g. following distances). Along with that it is recommended to also sample data on positive emotional reactions when driving with the assistant, for example when experiencing the benefits of catching green lights by sticking to the recommendations.

Acknowledgements

The research was conducted in the research project UR:BAN Urbaner Raum: Benutzergerechte Assistenzsysteme und Netzmanagement funded by the German Federal Ministry of Economics and Technology (BMWi) in the frame of the third traffic research program of the German government.

References

Deffenbacher, J.L., Lynch, R. S., Oetting, E.R., & Swaim, R.C. (2002). The Driving Anger Expression Inventory: A measure of how people express their anger on the road. *Behaviour research and therapy, 40*, 717-737.

Guéguen, N., Meineri, S., Martin, A., Charron, C. (2014). Car status as an inhibitor of passing responses to a low-speed frustrator. *Transportation Research Part F: Traffic Psychology and Behaviour, 22*, 245-248.

Maag, C. (2013). Emerging Phenomena During Driving Interactions. In E. Mitleton-Kelly (Ed.), *Co-evolution of Intelligent Socio-technical Systems: Modelling and Applications in Large Scale Emergency and Transport Domains* (pp. 185-218), Berlin Heidelberg, Springer.

Shinar, D. (1998). Aggressive driving: the contribution of the drivers and the situation. *Transportation Research Part F: Traffic Psychology and Behaviour, 1*, 137-160.

Stephens, A.N., Groeger, J.A. (2014). Following slower drivers: Lead driver status moderates driver's anger and behavioural responses and exonerates culpability. *Transportation Research Part F: Traffic Psychology and Behaviour, 22*, 140-149.

Tango, F., & Montanari, R. (2006). Shaping the drivers' interaction: how the new vehicle systems match the technological requirements and the human needs. *Cognition, Technology & Work, 8*, 215-226.

Thoma, S., Lindberg, T., & Klinker, G. (2007). Speed recommendations during traffic light approach: a comparison of different display concepts. In D. de Waard, F.O. Flemisch, B. Lorenz, H. Oberheid, and K.A. Brookhuis (Eds.), *Human Factors for assistance and automation* (pp. 63-73). Maastricht, the Netherlands: Shaker Publishing.

Tielert, T., Killat, M., Hartenstein, H., Luz, R., Hausberger, S., & Benz, T. (2010). The impact of traffic-light-to-vehicle communication on fuel consumption and emissions. In *Proceedings of the 2nd Internation Conference of Internet of Things (IoT),* (pp.1-8). IEEE.

Olfaction influences affect and cognitive-motoric performance: Evidence for the negative impact of unpleasant odours

Stefan Brandenburg, Anna K. Trapp, Nils Backhaus
Technische Universität Berlin
Cognitive Psychology and Cognitive Ergonomics
Germany

Abstract

Odours have been shown to affect mood as well as cognitive abilities. In this line of work specific odours like lavender, peppermint or ylang ylang have been examined. This paper examines whether the pleasantness of odours has an impact on participants affect and cognitive-motoric performance. Therefore a preliminary study was conducted in which 24 adolescent participants were exposed to either pleasant (e.g. pine tree) or unpleasant odours (e.g. soaked smoked cigarettes). Before and after being exposed to either one of both odours, subjects rated their affective status and completed the lane change task. Results showed that the interindividual experience of pleasantness differs much more for pleasant odours than for unpleasant odours. Futhermore participants felt significantly less positive and showed decreased lane change performance after being exposed to unpleasant odour, while pleasant odours showed no such effects. It can be concluded that unpleasant odours induced negative affect and influenced subjects' performance in this cognitive-motoric task. A possible application of these results could be the driving context where sensory input is one of the main factors for longitudinal and lateral vehicle control. In addition to visual, acoustic and tactile information, olfactory stimuli could also influence driving. However, subsequent studies should address real drivers in realistic driving scenarios.

Introduction

Most people do not doubt the importance of hearing and vision in their lives, but it is uncommon to think about the sense of smell as influencing ones behaviour and experience (Wrzesniewski, McCauley, & Rozin, 1999). Yet the olfactory system is closely associated with the limbic system (Sugawara et al., 2013) and odours modulate affect, behaviour, autonomic parameters and cerebral activity (Pollatos et al., 2007). In detail the piriform cortex and the amygdala are structures constituting the primary olfactory cortex while the insula and the orbitofrontal cortex belong to the secondary olfactory cortices (Doty et al., 1997; cited in Pollatos et al., 2007). This close physiological relationship between the olfactory system and the limbic system strengthens the hypothesis that odours stimulate positive and negative affect. Thereby odour research needs to consider dispositional preferences that can be aquired on an individual or culturally shared level (cf. Desmet & Heckert, 2007).

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

For example gaseous emissions in agriculture can constitute health problems for exposed workers, and odours from livestock affect the well beeing of nearby residents (Nimmermark, 2004). But odours do not only impact people's affect. They also manipulate information processing behaviour. Moss, Cook, Wesnes and Duckett (2003) showed that both lavender and rosemary caused a significant impairment of speed of memory. Lavender decreased the quality of working memory while rosemary enhanced its overall quality. Peppermint odour has also been argued to enhance memory (Moss et al., 2008). In terms of behaviour odours cause actions of approach or avoidance, at simplest. In a visual-tactile dual-task scenario, Ho and Spence (2005) showed a positive effect of peppermint odour on performance. Subjects reacted faster in a vibro-tactile task if exposed to peppermint scent.

Summing up, olfactory information influences a wide range of people's affect and behaviour. Previous studies showed that odours influence their well being, information processing and behaviour. To date few experimental investigations looked at the effect of odours on subjects' affect and behaviour in cognitive motoric tasks. The present pilot study aims at generating first indications whether it is worth exploring the role of odour in this type of tasks.

Objectives

The present investigation has two objectives. First, it examines the influence of pleasant and unpleasant odours on participants' affect. As the sense of smell is closely related to emotions, positive odours should elicit a positive affect and negative odours should lead to a negative affect (e.g. Pollatos et al., 2007). Second, the study investigates the effect of pleasant and unpleasant odours on a cognitive-motoric task. Regarding this question, positive odours (pine tree, perfume) should relax participants (Berneker, 2008) and therefore should increase their performance in the lane change task. Negative odours (soaked smoked cigarettes or acetone) should, in contrast, distract participants and lead to a reduced performance.

On these accounts, a pre-study was conducted to differentiate between pleasant and unpleasant odours. In the main study, subjects were exposed to either pleasant or unpleasant odours followed by a standardized cognitive-motoric task, the lane change task.

Method

Pre-study

The pilot study was conducted to distinguish between pleasant and unpleasant odours. For that reason 13 subjects (2 male) with an averaged age of M = 37.5 (ranging from 14 – 76 years) rated 10 everyday odours in a randomised order. All odours (jasmine, strawberry, pine tree, perfume, christmas mix, vinegar-based cleaner, chlorine, soaked smoked cigarettes, acetone, petrol) were presented in liquid form (1.2 ml) in opaque bottles. Subjects task was to open the bottle, smell the odour for about 30 s and rate their experience with respect to pleasantness (9-point Likert-type scale) and intensity (7-point Likert-type scale) following a standardized

procedure (see Sucker, Bischoff, Krämer, Kühner & Winneke, 2003 for more details). Results showed that participants rated the smells of jasmine, pine tree and perfume as most pleasant and cigarette, acetone and petrol as most unpleasant. There was no difference in intensity between odours.

Main study

Subjects
Twenty-four adolescents (16 male) with an average age of 13.74 years (*SD*=0.41) were tested in the main study. Almost all of them had prior experience with driving games and 42% of the subjects stated to play frequently.

Material
The independent variable pleasantness of odour was varied between subjects on the levels pleasant and unpleasant odours. For that reason, the two most pleasant (pine tree and perfume) and two most unpleasant (soaked smoked cigarettes and acetone) odours from the pilot study were used in liquid form (1.2 ml each) filled in opaque bottles for manipulating subjects' affect and performance in the main experiment. A short questionnaire was used to evaluate the subjective experience of pleasantness and intensity. This questionnaire was the same as in the pre-study and consisted of a bipolar item for pleasantness (9 point rating) and a bipolar item for intensity (7 point rating, Sucker et al., 2003, p. 26).

To measure affect a German version of the affect grid was used. The affect grid (Russel, Weiss, & Mendelsohn, 1989) consists of a 9x9 grid with the two-axis valence (extremely negative – extremely positive) and arousal (extremely sleepy – extremely aroused). Its theoretical basis is the circumplex model (Russel, 1980). The lane change task was used as the cognitive-motoric task. It is a standardized driving simulation that was shown in parallel on four desktop PCs with a 19-inch screen each. Subjects could change the speed and steering via the arrow keys on standard keyboards. Maximum speed was set at 60 km/h. Participants heard the simulator sound via earphones.

Procedure
Both conditions (positive and negative odours) were tested in two separate rooms with two subjects in each room at a time. After entering the room, participants were separately placed in front of a PC. Now they received a short introduction into the goals and the course of the experiment, the questionnaires and the lane change task. After that, they completed the first affect grid. Subsequently they had three minutes time to complete the practice trial of the lane change task. They were instructed to hold the speed at its maximum of 60 km/h at all times. Moreover they should change the lane as early as possible. Following the practice trial, subjects had the opportunity to ask questions. Now they performed another 3-minute section of the lane change task. These data were used as baseline. Another affect grid and the odours followed. As for the pilot study, the two positive or the two negative odours were presented in liquid form in opague bottles to each subject and participants were instructed to hold one bottle at a time directly under their noses and smell it for 30s. Subsequently to each smelling participants rated their subjective experience of the odour and their affective mood. They closed the lids of the bottles and accomplished

the test track of the lane change task. Again, this track consisted of a 3-minute stretch. Before starting this section, the experimenter refreshed participants' instruction to keep the speed at 60km/h and change the lane as soon and as quickly as possible. Due to the high intensity of the odours and the long smelling interval, the scent of the odours stayed in the room during the ratings and the test drive. After each group of participants the room was thoroughly aired. A short debriefing followed after the final test track. The experiment was part of a larger set of studies.

Results

To insure that the odours were experienced as pleasant and unpleasant, the ratings for pleasantness were evaluated (Table 1). All odour ratings' means were significantly different from zero, except for the pine tree. To sum up, the manipulation for pleasant odours was only partly effective while the manipulation for unpleasant odours was successful.

Table 1. Means, SD and one sample t-test against zero for pleasantness ratings of the four odours

	Mean Pleasantness	*SD Pleasantness*	*One sample t-test*
Pine tree	0.2	2.0	t(10) = 0.311 , NS
Perfume	1.8	1.6	t(11) = 3.783 , p = 0.003
Soaked smoked cigarettes	-2.1	2.2	t(11) = -3,354 , p = 0.006
Aceton	-1.3	1.5	t(11) = -3,084 , p = 0.010

Note. NS = non significant.

For analyzing whether odours affected subjects' affect, difference values for the affect grid scores of the baseline (without odours) and the test condition (with odours) were computed. The same applies for the question if odours affect behaviour. Here performance measures of the test condition (mean and standard deviation of the lateral position in the lane change task) were subtracted from the scores of the baseline condition. Due to a setting error six subjects had a smaller viewing distance in the lane change task. To compensate the difference the computational model of the standard line was adjusted. Both analysed measures were not affected by this, neither the mean deviation of lateral position (t(22)=-0.317, NS) nor the standard deviation of the lateral position (t(22)=-0.539, NS).

Effect of odours on affect

With respect to the effect of odours on subjects' affect, no effect was found for unpleasant and pleasant odours on arousal, all $t < 0.37$, all $p > 0.71$. In contrast, unpleasant odours significantly decreased subjects valence scores, $t(11) = 4.7$, $p <$

0.001. Participants of this group felt less positive after being exposed to unpleasant smells. For pleasant odours, no effect on subjects valence ratings was obtained, t(11) = 0, NS. Figure 1 visualizes the results.

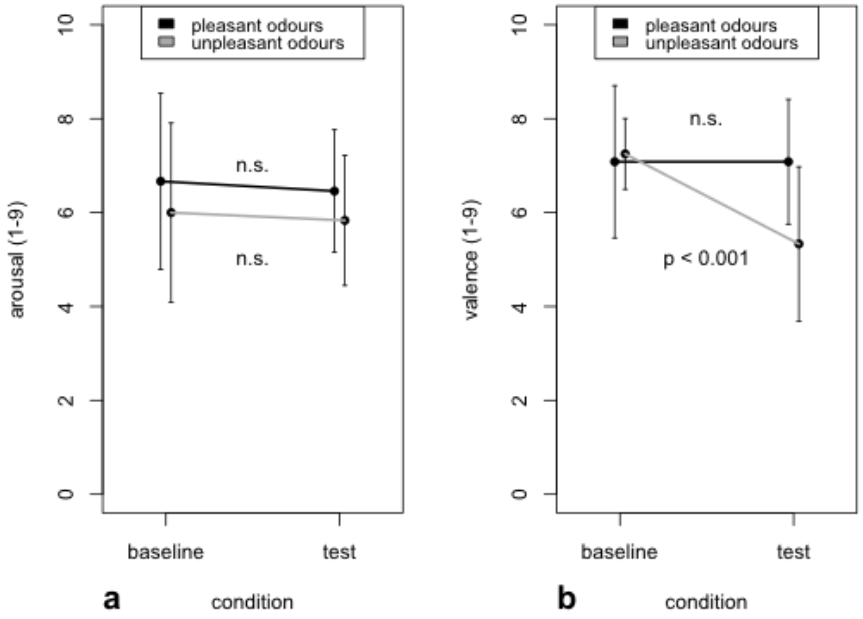


Figure 1. The effect of odours on subjects a) arousal and b) valence. Error bars represent +/- 1SD; baseline = before odour exposure, test = after odour exposure.

Effect of odours on the performance in the cognitive-motoric task

With respect to the effect of odours on cognitive-motoric performance, effects of odours on the mean lateral position and the standard deviation of the lateral position were found. Subjects that were exposed to pleasant odours showed a tendency with respect to a decreased lateral deviation compared to their baseline, t(10)=1.49[1], p = 0.08. In contrast, unpleasant odours resulted into a tendency for an increased lateral deviation, t(11) = -1.63, p = 0.06. Moreover, pleasant odours did not affect the standard deviation of the lateral position, t(11) = -0.13, p = 0.55. Again unpleasant odours increased the standard deviation of the lateral position, t(11) = -1.88, p = 0.04. Figure 2 visualizes the effects of odours on a) the mean deviation from the lateral position and b) the standard deviation of the lateral position.

[1] One subject was excluded from the group of pleasant odours because of being a large outlier.

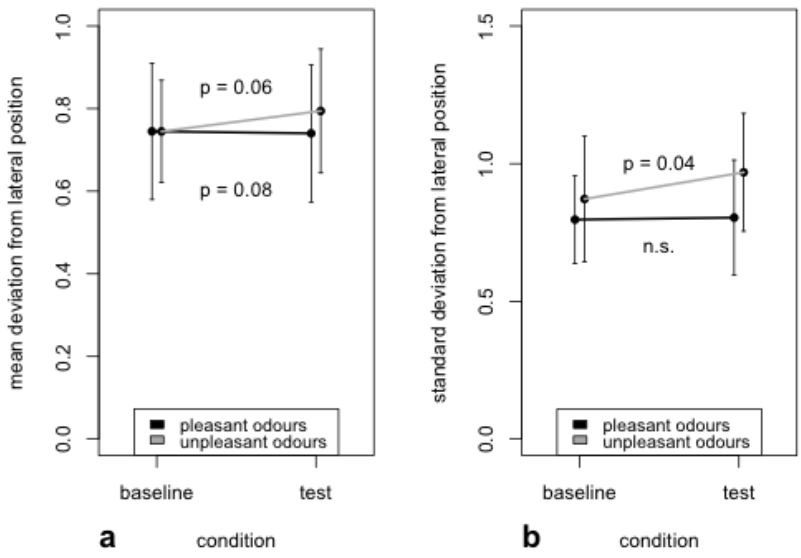


Figure 2. Effect of odours on a) the mean deviation from the lateral position and b) the standard deviation of the lateral position. Error bars represent +/- 1SD; baseline = before odour exposure, test = after odour exposure.

Discussion

The present study had two objectives. First, it examined whether everyday odours affect subjects' affect. Regarding this objective, only one significant difference was obtained for participants' valence ratings after they had been exposed to negative odours. This result somewhat deviates from literature findings that used standardized odour samples like the sniffing sticks (e.g. Pollatos et al., 2007). One explanation for this result lies in the fact that participants experienced only one of the two positive odours as pleasant while both negative samples were experienced as unpleasant. Thus the manipulation for pleasant odours was not successful. Kaye (2004) points out several issues when designing pleasant and unpleasant odours of which probably the main problem is interindividual variance in odour perception and judgement. Even though the odour samples were choosen based on a pre-study, two different samples with a different range of age participated in the pre-study and the main study. The difference in sample characteristics might explain these findings partially. Future experiments should try to individually determine pleasant and unpleasant odours or use within-subjects designs with the same subjects in the pre- and the main study. Using personalized stimuli or a different experimental design, a replication of effects from literature with everyday odours might be more likely.

Second, the present study investigated whether pleasant and unpleasant odours affect performance in the lane change task, a simple cognitive-motoric task. When operating this simulation participants have to continuously adjust their lateral position based on visual input. Results indicate that subjects showed a tendency towards better steering performance in this cognitive-motoric task when being

exposed to pleasant odours. Moreover a tendency for worse steering performance was shown for the negative odour group. However, these hypothesis confirming result were just tendencies and only applied to the mean deviation of the simulated vehicles lateral position. In contrast the significant effect of negative odours on the variability of the cognitive-motoric performance seems to be trustworthier. Performance decreased when participants were previously exposed to unpleasant odours. This was shown in the marginal increase of the mean and the significant increase of the standard deviation of the lateral position. Subjects might have been distracted by the unpleasant smell. For example Wrzesniewski et al. (1999) argue that subjects feel the urge to avoid or seek out unpleasant smells. This behavioural tendency even increases with increasing unpleasantness of odours. Therefore participants might have concentrated on their breathing or other strategies of avoiding unpleasant smells instead of concentrating on the cognitive-motoric task. Pleasant odours, in contrast, foster the subject to increase their experience of them (Wrzesniewski et al., 1999). Thus, subjects that were exposed to positive odours were not distracted and could concentrate on the cognitive-motoric task. This explanation seems reasonable since the smell of the odours stayed in the room even after the active smelling and was only removed after the test track.

Summing up, the present study showed that affective states and keeping lateral control in a simple driving simulation was affected by everyday odours. Thus we conclude that the dimension pleasantness of odour indeed has an impact on affect and cognitive-motoric performance of adolescents. Nevertheless the conclusion is limited to the specific sample and to only one pole of pleasantness since the manipulation for pleasant odours was only partial successful. A practical application of this study could be the context of car driving. Here, having longitudinal and lateral control over the vehicle, both cognitive-motoric tasks, is extremely safety relevant. While studies in this field mainly focused on visual, acoustic, tactile modalities and higher cognitive factors, olfactory stimuli could also influence driving performance. This study is a small but relevant step towards more applied research on the olfactory influences on subjects' affect and behaviour in human-machine interaction situations.

Acknowledgement

We would like to offer our special thanks to Ron Reckin and Thorsten Fischer for spending their valuable time on the project during data collection and to Dick de Waard for his helpful comments.

References

Berneker, A. (2008). Der Duft der Autos. *Fahrschule*, *12*, 3-22.

Desmet, P. M. A., & Hekkert, P. (2007). Framework of product experience. *International Journal of Design, 1*(1), 57-66.

Ho, C., & Spence, C. (2005). Olfactory facilitation of dual-task performance. *Neuroscience Letters*, *389*(1), 35–40.

Kaye, J. (2004). Making scents: Aromatic output for HCI. *Interactions*, 49-61.

Moss, M., Cook, J., Wesnes, K., & Duckett, P. (2003). Aromas of rosemary and lavender essential oils differentially affect cognition and mood in healthy adults. *International Journal of Neuroscience, 113*, 15–38.

Moss, M., Hewitt, S., Moss, L., & Wesnes, K. (2008). Modulation of cognitive performance and mood by aromas of peppermint and ylang-ylang. *International Journal of Neuroscience*, *118*(1), 59–77.

Nimmermark, S. (2004). Odour influence on well being and health with specific focus on animal production emissions. *Ann Agric Environ Med 11*, 163-173.

Pollatos, O., Kopietz, R., Linn, J., Albrecht, J., Sakar, V., Anzinger, A., Schandry, R., & Wiesmann, M. (2007). Emotional stimulation alters olfactory sensitivity and odor judgment. *Chemical Senses*, 1-7.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*, 493–502.

Sucker, K., Bischoff, M., Krämer, U., Kühner, D., & Winneke, G. (2003). *Forschungsbericht: Untersuchungen zur Auswirkung von Intensität und hedonischer Geruchsqualität auf die Ausprägung der Geruchsbelästigung*. Medizinisches Institut für Umwelthygiene an der Heinrich-Heine Universität Düsseldorf.

Sugawara, Y., Shigeto, A., Yoneda, M., Tuchiya, T., Matumura, T., & Hirano, M. (2013). Relationship between mood change, odour and its physiological effects in humans while inhaling the fragrances of essential oils as well as linalool and its enantiomers. *Molecules 18*, 3312-3338.

Wrzesniewski, A, McCauley, C & Rozin, P. (1999) Odor and affect: Individual differences in the impact of odor on liking for places, things and people. *Chemical Senses 24*, 713-721.

The more the better? The impact of number of stages of likelihood alarm systems on human performance

Magali Balaud & Dietrich Manzey
Technische Universität Berlin
Germany

Abstract

Responses to alarms involve decisions under uncertainty. Operators do not know if an alarm is more likely to be a hit or a false alarm. Likelihood alarm systems (LAS) help reduce this uncertainty by providing information about the certainty of their output. Unlike traditional binary alarm systems, they have three or more stages: each one represents a different degree of likelihood that a critical event is really present. Consequently, the more stages, the more specific is the information provided by the alarm system to reduce uncertainty. A laboratory experiment with 48 participants was conducted to investigate the effect of specificity of information of LAS on performances and responding behaviour. Specifically, a three-stage, four-stage, and five-stage LAS were compared using a multi-task environment. Results show higher percentages of correct decisions in the alarm task when participants used the four- and five-stage LAS than the three-stage LAS but no significant differences were found between the four-and five-stage LAS. Interesting differences in response patterns were also observed. This study suggests that four stages is the best degree of specificity for optimal performance.

Introduction

Alarm systems are extremely useful in multitasking and high workload environments such as aviation cockpits, hospitals and industries. They play a role of mediator between a human operator and a process, receiving information about the current status of a process and informing operators about it so that critical events are not missed. Most of the time operators work with Binary Alarm Systems (BAS) which inform the operator in a binary way: there is a critical event (red) or not (green).

Ideally, an alarm should go off only if there is a critical event. However this is not always the case. Instead alarms systems usually tend to generate a lot of false alarms, i.e. alarms go off even if there is no critical event. This is partly due to the "engineering fail-safe approach" (Swets, 1992): in order not to miss any critical events, engineers design the alarm system so it goes off even if there is little evidence of a critical event. A useful descriptor of the reliability of an alarm system is the Predictive Positive Value (PPV) (Getty et al., 1995). The PPV is the conditional probability that, given an alarm, a problem actually exists. A PPV of 0.3, e.g., means that out of all alarms emitted by the system, 30% are hits and 70% are false alarms. Given that alarm systems in most domains emit a high number of false alarms their PPV is usually low, often less than 0.1 (Parasuraman & Riley, 1997).

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

As a consequence, operators might stop trusting them (Madhavan, Wiegmann & Lacson, 2006). In behavioural terms, this can lead to what has been referred to as the cry-wolf effect (Breznitz, 1984; Getty et al., 1995; Bliss et al., 1995). Operators tend to respond slower or even to ignore the alarm system when it goes off. This can result in dramatic consequences regarding the safety and productivity (Lee & See, 2004).

One possible solution to address this issue is the use of Likelihood Alarms Systems (LAS). This concept was first developed by Sorkin et al. (1988) to constitute an alternative to binary alarm systems. LAS are composed by three or more stages with each stage corresponding to a different likelihood that a critical event is present. In other words, each stage of LAS has a different PPV and communicates it to the operator through the use of different colours, wordings, or sounds.

The goal of LAS is to provide more differentiated information to operators than traditional binary alarm systems so that they can adapt their responding behaviour depending on how likely it is that a critical event is present. By adapting their responding behaviour properly to the PPV of each stage, operators have higher chances to correctly comply with hits and to correctly ignore false alarms produced by the alarm system. Previous laboratory studies have shown that participants respond less to LAS in comparison to BAS but that they are more accurate: operators produce more hits and fewer false alarms with LAS in comparison to BAS (Bustamante & Bliss, 2005; Wiczorek & Manzey, 2014).

This raises the question of what degree of specificity, i.e. number of stages of LAS, is optimal for operators. Two studies (Shurtleff, 1991; Wiczorek et al., 2014) have already investigated this question. Shurtleff compared a BAS, a 4-stage LAS, a 6-stage LAS, an 8-stage LAS, and a control condition in which participants did not get any advice from any alarm system. The difficulty of the decision task was also manipulated. Results show that only when the task is difficult does the number of stages on participant's performance have an effect. Participants showed better performance while using 4-stage LAS and 8-stage LAS than BAS or no alarm. Wiczorek et al. (2014) compared a BAS, a 3-stage LAS, and a 4-stage LAS supporting a monitoring task as part of a multi-task scenario. They found that participants made less incorrect decisions (i.e., misses and false alarms) when they used the 4-stage LAS, followed by the 3-stage LAS and the BAS.

The current study

The current study investigates the optimal number of stages in Likelihood Alarm Systems on participants' responding behaviour, participants' performance and participants' workload. Using the same task environment than Wiczorek et al. (2014), the aim of this study was to replicate their findings using different PPV alarm characteristics and to further investigate the question of the optimal number of stages in LAS by comparing a 3-stage, 4-stage, and 5-stage LAS. The 3-stage LAS was composed by a non-alarm stage, a warning stage, and an alarm stage. Based on that, the 4-stage LAS was created by dividing the warning stage in two stages while the alarm stage was kept constant. The same logic applied in order to make the 5-

stage LAS: the stage of the 4-stage LAS having the lowest PPV (i.e., the yellow-warning stage) was split into two stages.

The following hypotheses were addressed: Firstly, it was hypothesized that participants would adapt their responding behaviour to the PPV of each stage so that participant's response rate in each stage will significantly be different from the others. Secondly, a differentiation in participants' behaviour would be expected between the 3-stage LAS and the 4-stage LAS. Specifically, it was assumed that the cry-wolf effect would be shifted from the warning stage of the 3-stage LAS to the low-PPV warning stage of 4-stage LAS and that participants would comply more with the high-PPV warning stage of the 4-stage LAS than with the warning stage of the 3-stage LAS. A similar effect was expected between the 4- and 5-stage LAS.

Thirdly, regarding participants' performance in the alarm task, a main effect of the number of stages on participants' decision-making performance was expected. The more stages, the better participants' performance would be in terms of the percentage of hits and false alarms. More specifically, participants' percentage of hits would increase with the number of stages and participants' percentage of false alarms would decrease with the number of stages.

Fourthly, with respect to participants' performance in the concurrent tasks, a decrease of performance was expected in the 5-stage LAS condition only. As too much specificity (stages) in the alarm display might increase the workload and time-demands of decision-making in response to the alarm system, it was assumed that increasing specificity might negatively impact operators' ability to deal with concurrent tasks. Since Wiczorek et al. (2014) did not find any difference between the 3-stage LAS and the 4-stage LAS on concurrent tasks performance, a visible decrease of performance was expected only for the most complex 5-stage LAS. Finally, it was expected that the more stages the LAS have, the higher participants' workload would be.

In addition to the hypotheses-driven questions, participants' overall response rate towards alerts (i.e., alarms and warning together) was also investigated in an exploratory manner, in order to know to what extent the number of stages of LAS would impact the cry-wolf effect.

Method

Participants

Forty-eight participants (22 men, 26 women) participated in this study. Participants ranged in age from 18 to 44 years with a mean age of 27.02 years (SD = 5.77). None of them was suffering from any distortion of colour vision which might interfere with the experiment (i.e. red-green colour blindness). Participants were paid 5€ for their participation and they could get an additional bonus of maximum 4€ depending on their performance during the experiment.

Task

The PC-based Multi-Task Operator Performance Simulation (M-TOPS) was used. It simulates in a simplified way typical multi-task demands of operators in a control room. Participants had to accomplish three tasks simultaneously. In one of these tasks, they were assisted by an alarm system. A picture of the M-TOPS interface is shown in Figure 1.

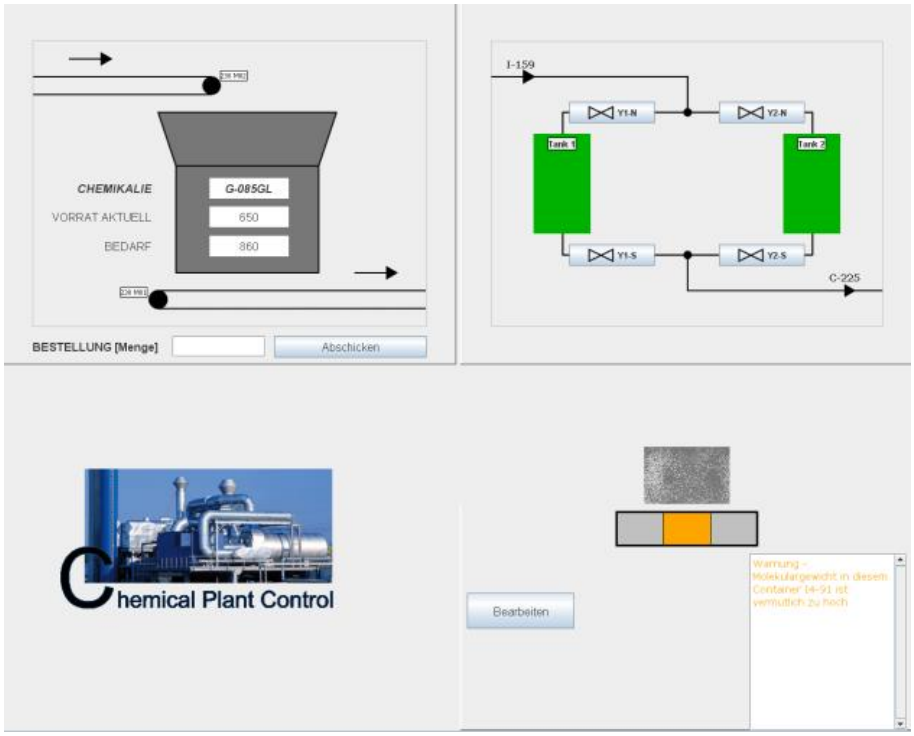


Figure 1. User interface of M-TOPS

Resource Ordering Task (ROT). This task is a mental arithmetic task displayed in the upper left quadrant of the interface. Participants are instructed that they have to ensure the availability of required chemicals in order to keep the chemical process running. For this purpose, the current and the required value of a chemical is presented. Participants are asked to calculate the arithmetic difference, type the result in the designated ordering field, and initiate the order by clicking a button. They received 1.5 cents for each correctly sent order.

Coolant Exchange Task (CET). This task is displayed on the upper right quadrant of the interface. Participants are responsible for exchanging the coolant in different sub-systems of the plant. To do this they have to open and close a few valves by clicking on them following a certain order. A complete exchange cycle takes about 40 seconds. Participants received 7.5 cents for each refilling cycle successfully completed.

Alarm Task (AT). In this task displayed in the lower right quadrant of the interface, participants have to decide if the final quality of the chemical product has a correct molecular weight. They are assisted by an LAS showing a different *colours* and wordings depending on how likely it is that the chemical product has an improper molecular weight. Based on the diagnostic of the LAS participants choose between sending the container back to the plant (by clicking on the repair button) or letting it go (by doing nothing). Participants have no other cues apart from the output of the alarm system to help them in their decision. They lose 2 cents for each wrong decisions (i.e., repairing a correct container or ignoring an improper container). This pay-off was chosen based on a precise analysis of how much time participants spend on each task. It aims to keep a constant competition between the different tasks so that no task is left out for strategic reasons. The same pay-off was also used in the works of Wiczorek & Manzey (2014) and Wiczorek et al. (2014).

Design and alarm systems characteristics

The experimental design was composed of a single between-subjects factor defined by the number of stages of the likelihood alarm system supporting the alarm task. This factor had three levels: 3-stage (LAS3), 4-stage (LAS4), and 5-stage (LAS5). All alarm systems had the same sensitivity (d = 1.8). The basic characteristics of the three alarm systems used are presented in Figure 2. The first criterion separating the non-alarm stage ("green") from the other stages was kept constant for all systems (c = -1.05). The numbers reported in the squares correspond to the PPV of each stage and the number reported under each separation corresponds to the criterion. The colours presented in this figure are the colours used for the outputs of the LAS. They were chosen according to findings from previous studies investigating the link between colours and perceived urgency or perceived hazard (Braun & Silver, 1995; Chapanis, 1994; Wolgater et al., 2002).

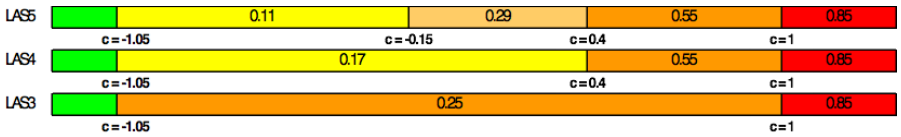


Figure 2. Systems characteristics of the three LAS

Dependent variables

Alarm task response behaviour: Possible differences of participants' responses to the different stages of the different LAS were assessed by their compliance rates with each stage. Compliance rate was defined as the percentage of alerts emitted by each stage which was responded to by a click on the repair button.

Alarm task performance: Participants' performance in the alarm task was assessed by the average percentage of hits and false alarms achieved by the participants in interaction with the different LAS. A high percentage of hits as well as a low percentage of false alarms is considered as good performance.

Concurrent tasks performance: Participants' performance in the concurrent tasks was measured by the amount of correctly sent orders in the Resource Ordering Task (ROT) and the amount of refilling cycles successfully completed in the Coolant Exchange Task (CET).

Subjective workload: Participants' perceived workload was assessed using the NASA Task Load Index (Hart & Staveland, 1988). The mean of all six single scales was considered as overall workload measure.

Procedure

Participants first completed an informed consent form and a demographic questionnaire and were then provided with the task instructions on the computer screen. They were told that the experiment was a simulation of a control room of a chemical plant and that they had to perform three tasks concurrently in order to assure the good run of the chemical process and to control the quality of the end-product. Participants had a 2-minute training for each single task. They were then explained that the alarm system was not 100% reliable and that it could sometime provide wrong outputs. This was followed by a 50-trial familiarization session (about 8 minutes) in which participants performed the alarm task only and received an auditory feedback after each decision in response to the outputs of the alarm system they made. The feedback informed them about the correctness of their decision and, thus, implicitly also about the performances of the alarm system. They were told to use this auditory information to get an idea of the reliability of the different stages of the LAS. Participants were then explicitly asked for a subjective assessment of the reliability of each stage of the LAS they had worked with. This was used as a manipulation check to ensure that participants paid attention to the auditory feedbacks in the familiarization session and recognized the differences in PPVs of the different stages. The experimental session finally started. It was composed of 100 containers (about 16 minutes). No auditory feedbacks were provided during this session. Finally participants completed the NASA TLX questionnaire, were thanked for their participation and received a monetary compensation.

Results

Participants' response behaviour

Response rates for the 3-stage LAS

Response rates to the two alert stages of the LAS3 (alarm vs. orange-warning) are shown in Figure 3. As expected participants on average complied more with alarms (98.56%) than warnings (16.51%). This difference was proven to be statistically significant by a two-tailed t-test, $F(1,15) = 120.58$, $p = .000$.

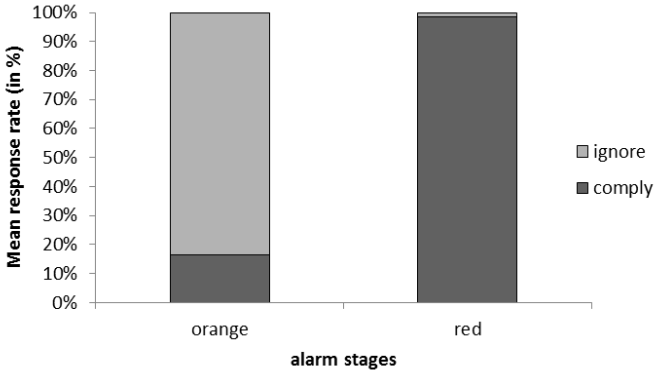


Figure 3. Means of compliance rates and non-compliance rates towards the 3-stage LAS depending on the diagnosis emitted by this LAS.

Response rates for the different stages of LAS4

Mean response rates for the three alert stages of the LAS4 (alarm vs. orange-warning vs. yellow-warning) are displayed in Figure 4. As becomes evident, response rates differed between stages. A one-way ANOVA with stage (red-alarm, orange-warning, yellow-warning) as within factor was used to analyse this effect. This was composed by a linear contrast C1(-1, 0, 1) and a quadratic contrast C2 (-1, 2, -1). The linear contrast was significant suggesting that participants complied more with alarms than yellow-warnings, $F(1, 15) = 111.68$, $p = .00$. However the quadratic trend was also significant showing that participants' compliance rate towards orange-warnings differed from the linear trend, $F(1, 15) = 111.03$, $p = .00$. The significance of the quadratic trend is explained by the high compliance rate observed with orange-warnings (97.16%), which does not significantly differ from participants' compliance rate with alarms (96.63%), $F(1, 15) = .10$, $p = .76$.

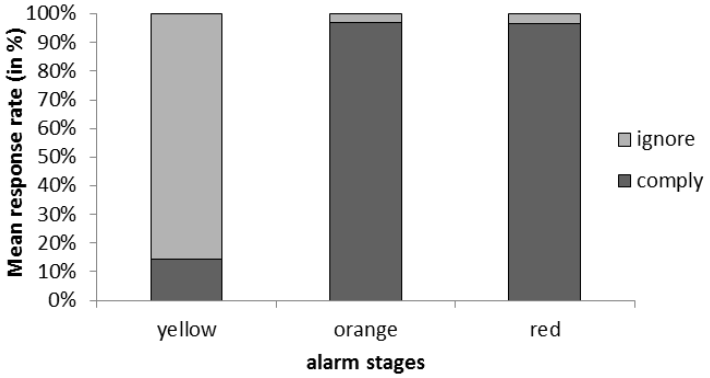


Figure 4. Means of compliance and non-compliance rates toward the 4-stage LAS depending on the diagnosis emitted by this LAS.

Response rates towards different stages of LAS5
Results are displayed in Figure 5. A one-way ANOVA with stage (alarm vs. orange-warning vs. orange-yellow-warning vs. yellow-warning) as within factor was used for the analysis of the response rate toward LAS5. A linear contrast C1 (-3, -1, 1, 3), a quadratic contrast C2 (-1, -1, 1, -1) and a cubic contrast (-1, 3, -3, 1) were used to test how specifically participants' responses to the different stages depends on the PPV of each stage. The linear trend is significant, $F(1, 15) = 120.34$, $p = .00$, as well as the cubic trend, $F(1, 15) = 5.31$, $p = .04$. This means that the pattern of results is not completely linear as expected. The high compliance rate obtained in the orange-warning stage is responsible for the significance of the cubic trend. This was confirmed by the fact that participants' compliance rate did not differ in the orange-warning stage and the red-alarm stage, $F(1, 15) = 2.46$, $p = .14$.

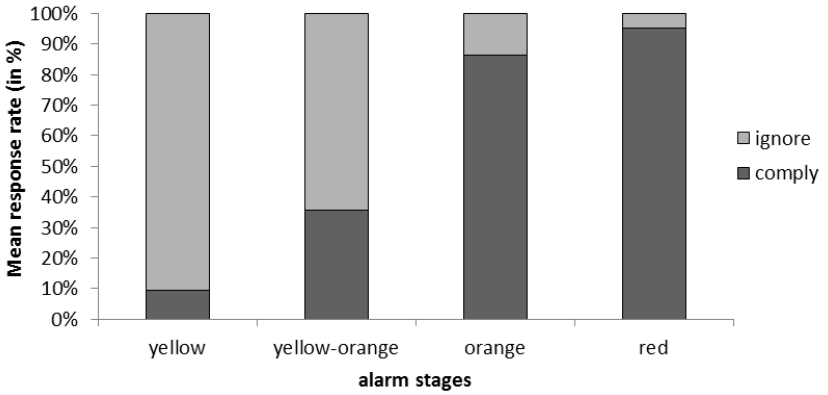


Figure 5. Means of compliance and non-compliance rates toward the 5-stage LAS depending on the diagnosis emitted by this LAS.

Comparisons of response behaviour across different LAS
A one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor was used for the analysis of the response rate toward alerts. Even though participants complied more with LAS4 (44.51%) and LAS5 (44.70%) than with LAS3 (32.67%), these difference were not significant, $F(2, 45) = 1.7$, $p = .19$. This means that the cry-wolf effect, in terms of number of percentage of ignored alerts, was the same among the three LAS.

However a behavioural differentiation was observed as expected in Hypothesis 2. Participants complied significantly more with the orange warning stage of LAS4 (97.16%) than with the warning stage of LAS3 (16.51%), $F(1, 30) = 107.91$, $p = .00$. Moreover, participants complied significantly more with orange warnings of LAS4 than yellow warnings of LAS4, $F(1, 15) = 116.64$, $p = .00$, showing that the cry-wolf effect in LAS4 was reduced to the yellow-waning stage only. A shift of the cry-wolf effect from the warning stage of LAS3 to the yellow warning stage of LAS4 happened.

Regarding LAS4 and LAS5, participants did not significantly complied more with the yellow-orange warning stage of LAS5 (35.71%) than with the yellow warning

stage of LAS4 (14.58%), $F(1, 16) = 2.65$, $p = .11$, even though descriptive results show this tendency. A behavioural differentiation occurred still between the yellow warning stage and the yellow-orange warning stage of LAS5. Participants complied significantly more with the yellow-orange warning stage (35.71%) than the yellow warning stage (14.58%), $F(1, 16) = 7.27$, $p = .02$.

Alarm-Task performance

All analyses about participants' performance in the alarm task were performed using a one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor. Two orthogonal contrasts were defined for pairwise comparisons of means: C1 (2, -1, -1) and C2 (0; -1; 1). The first contrast C1 compares the mean performance for LAS3 with the combined mean performances for LAS4 and LAS5. The second contrast C2 tests if performances in conditions LAS4 and LAS5 would differ from each other.

Participants' percentage of hits and false alarms are displayed in Figure 6. Two participants were excluded from the analysis on the percentage of hits based on their outlying SDR and Cook values. One participant was excluded from the analysis on the percentage of false alarms for the same reasons.

Regarding the percentage of hits, results did not show a linear trend as it was

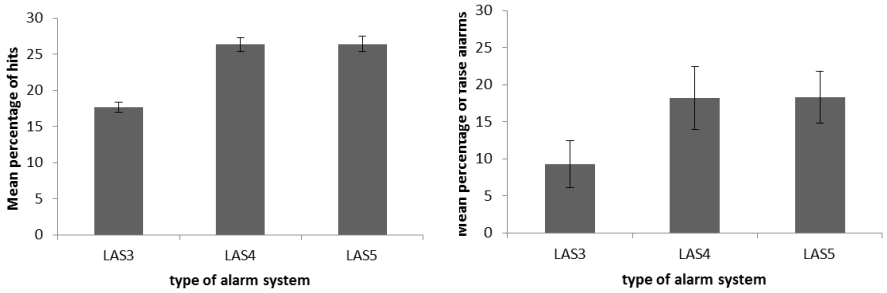


Figure 6. Means and mean standard deviations of participants' percentage of hits (left panel) and false alarms (right panel) in the alarm task depending on the type of LAS.

predicted. As expected, participants using LAS3 produced significantly less hits (17.64%) than participants using LAS4 (26.33%) but participants using LAS5 (26.42%) did not produce more hits than participants using LAS4. This pattern is also confirmed by the two contrasts, C1: $F(1, 44) = 52.91$, $p = .00$, C2: $F(1, 44) = 0.01$, $p = .94$ (C2).

Regarding participants' percentage of false alarms, the best performance (i.e., the lowest percentage of false alarms) was observed in the LAS3. Participants using LAS3 produced less false alarms (9.29%) than participants using LAS4 (18.18%) and LAS5 (18.27%), $F(1, 45) = 3.84$, $p = .05$ (C1). No difference between the LAS4 and LAS5 condition has been found, $F(1, 45) = 0.00$, $p = .99$ (C2).

Concurrent task performances

A one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor was used to analyse the performance data of the two concurrent tasks. No significant differences between the three conditions were found in both tasks: ROT: $F(2, 44) = 0.41$, $p = .66$; CET: $F(2, 45) = 0.06$, $p = .943$.

Workload

A one-way ANOVA with number of stages (3 vs. 4 vs. 5) as between factor was used for the analysis of the participants' workload. There is no main effect of number of stages on participants' workload ratings, $F(2, 45) = 1.05$, $p = .36$. No effect was found on any single scale of the NASA TLX.

Discussion

This study aimed to investigate what number of stages of likelihood alarm systems would provide the optimal specificity of information for human performance in interaction with such systems. Specifically, the effect of three different LAS on responding behaviour, performance and workload was investigated. The LAS differed with respect to the number of stages.

Participants adapted only partially their responding behaviour to the PPV of each stage. This means that the pattern of results is not exactly linear but shows a kind of dichotomization. Participants tend to clearly differentiate their responding behaviour depending on the PPV towards stages having a PPV under .5. This tendency of operators to adjust their response behaviour to the PPV of alerts at the lower end of PPVs was also reported by other studies addressing the impact of PPV on responses to alarms of BAS as well as studies investigating different stages of LAS (Manzey et al., 2014; Wiczorek & Manzey, 2014; Wiczorek et al., 2014). However participants tend to consistently comply with alerts emitted by stages having a PPV above .5. Participants complied with more than 93% of orange warnings emitted by the LAS4 and LAS5 even though the PPV is .55. This high compliance rate is actually a rational strategy in order to optimize the amount of correct decisions in interaction with alarm systems and is very surprising, as such high response rates are usually observed in stages having a PPV above .7 (Wickens & Dixon, 2007). Interestingly, adding more stages to LAS does not reduce the cry-wolf effect. However, while participants' overall response rate was the same for the three LAS, their overall decision-making performance in terms of hits clearly benefited from going from an LAS3 to an LAS4. By adding one more stage, thus providing more differentiated likelihood information, participants get more opportunities to differentiate their behaviour. The ignorance of alert, i.e. the cry-wolf effect, still occurs but is shifted to a stage having a lower PPV and thus shifted to a stage where an ignorance of the alert often matches an alert which is false anyway. As a consequence, participants comply more with true alarms and ignore more false alarms even though the overall response rate to alerts stays the same. Studies comparing BAS to LAS3 have even shown that participants' overall response rate is higher with BAS than LAS but performance is still better with the LAS which is attributed to essentially the same effect (Bustamante & Bliss, 2005; Manzey et al. 2014).

Regarding participants' performance in the alarm task, they showed better performance with the LAS4 and the LAS5 than the LAS3 with respect to the percentage of hits. However, no significant differences emerged between the LAS4 and LAS5. Against our expectations, participants had lower performance with the LAS4 and LAS5 than the LAS3 with respect to the percentage of false alarms. This is in contradiction with results reported by Wiczorek et al. (2014) showing that participants produce fewer false alarms with the LAS4 than the LAS3. The high response rate toward orange warnings in the LAS4 and LAS5 might explain these results. By complying with more than 93% of warnings having a PPV of .55, participants produced a great amount of false alarms in comparison to participants in the LAS3 condition who mainly ignored the .25 PPV warnings and produced mostly correct rejections. However the percentage of hits is a more relevant performance indicator to consider than the percentage of false alarms since most alarms systems are used in environment in which misses are more costly than false alarms. From these results, one can draw the conclusion that LAS4 improve performance over LAS3 and that adding one more stage (LAS5) does not improve performance further.

No effect of the number of stages in LAS has been found on participants' performance in the concurrent tasks. This is probably due to the fact that participants' workload did not increase with the greater amount of information provided by the LAS5. Indeed no difference between the three LAS on participants' workload has been found. It would be interesting, however, to know if a higher number of stages affect the workload since alarm systems having more than 5 stages are sometimes used in ecological environments.

Conclusion

Likelihood alarms systems are definitely an option to consider in situations in which the use of a BAS leads to a high cry-wolf effect with the performance effect of decreasing hit rates. This study suggests that a 4-stage LAS provides the optimal degree of specificity and that a higher degree of specificity does not improve performance. However, one limiting factor of the current research was that the participants did not get the opportunity to cross-check the validity of alarms before responding to it. Previous research has shown that providing such an option might significantly impact the response behaviour in interaction with alarms (e.g., Manzey et al., 2014). Further research is needed to investigate if the results reported in this study could be generalized to situations in which operators have access to alarm validity information.

References

Bliss, J., Dunn, M., & Fuller, B.S. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. *Perceptual and Motor Skills*, *80*, 1231–1242.

Braun, C.C., & Silver, N.C. (1995). Interaction of signal word and colour on warning labels: differences in perceived hazard and behavioural compliance. *Ergonomics*, *38*, 2207–2220.

Breznitz, S. (1984). *Cry Wolf: The Psychology of False Alarms*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bustamante, E.A., & Bliss, J.P. (2005). Effects of workload and likelihood information on human response to alarm signals. In *Proceedings of the 13th International Symposium on Aviation Psychology* (pp. 81–85). Oklahoma City, OK: Wright State University.

Chapanis, A. (1994). Hazards associated with three signal words and four colours on warning signs. *Ergonomics*, *37*, 265–275.

Getty, D.J., Swets, J.A., Pickett, R.M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied*, *1*, 19–33.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, *52*, 139-183.

Lee, J.D., & See, K.A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*, 50–80.

Madhavan, P., Wiegmann, D.A., & Lacson, F.C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors*, *48*, 241–256.

Manzey, D., Gérard, N., & Wiczorek, R. (2014). Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload. *Ergonomics*, *57*, 1833-1855.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*, 230–253.

Shurtleff, M.S. (1991). Effects of specificity of probability information on human performance in a signal detection task. *Ergonomics*, *34*, 469–486.

Sorkin, R.D., Kantowitz, B.H., & Kantowitz, S.C. (1988). Likelihood Alarm Displays. *Human Factors*, *30*, 445–459.

Swets, J.A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, *47*, 522–532.

Wickens, C.D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*, 201–212.

Wiczorek, R., & Manzey, D. (2014). Supporting Attention Allocation in Multitask Environments Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance. *Human Factors*, *56*, 1209-1221.

Wiczorek, R., Manzey, D., & Zirk, A. (2014). Benefits of Decision-Support by Likelihood versus Binary Alarm Systems: Does the number of stages make a difference? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*, 380-384. Santa Monica: HFES.

Wogalter, M.S., Conzola, V.C., & Smith-Jackson, T.L. (2002). Research-based guidelines for warning design and evaluation. *Applied Ergonomics*, *33*, 219–230.

The predictive quality of retentivity for skill acquisition and retention in a simulated process control task

Barbara Frank & Annette Kluge
Ruhr University Bochum
Germany

Abstract

Past studies have shown the potential of refresher interventions to mitigate skill decay in process control. More recent studies also indicate the predictive quality of retentivity as a person-related variable. The two presented studies investigated the impact of retentivity on non-routine tasks in the context of simulated ordinary work experience. Study 1 (*N*=18) compared four retentivity measures (Selective Reminding Test, WIT-2, I-S-T-2000R and Map Learning) as indicators of skill acquisition in a simulated process control task, and showed significant, moderate correlations between the target skill (production outcome) and Map Learning *directly* after training. Study 2 (*N*=39) investigated the retentivity constructs in the context of simulated work experience and skill retention, and consisted of four measurement times: 1.) initial training of the target skill (week 1), 2.) and 3.)work experience (target skill was not required; week 2 & following week 3) and 4.) the retention assessment of the target skill (week 4). The control group took part in initial training and retention assessment only. Results showed significant, moderate correlations between Map Learning and production outcome and between WIT-2 and production outcome in retention assessment (after the retention interval). Retentivity constructs and practical implications will be discussed based on these findings.

Introduction

The operator's tasks in highly automated plants such as in process control include monitoring the plant and its process, keeping records and adjusting the system (Kluge, 2014). In the case of emergency, however, if the plant is no longer controlled by the automated system, the operator has to make decisions and control the plant him/herself. In industries with a high level of automation, after long periods of non-use or in non-routine situations (defined above all by the rarity with which a particular skill is performed; Kluge, 2014), there is a particularly strong risk of decay of once learned skills and knowledge, meaning that the operator might not know what to do in an emergency (e.g. Bainbridge, 1983; Kaber, Omal, & Endsley, 1999; Parasuraman, Sheridan, & Wickens, 2000; Wickens & McCarley, 2008). Skill decay can be explained by the "Power Law of Forgetting" (Bourne & Healy, 2012) and the "New Theory of Disuse" (Bjork & Bjork, 1992; Bjork & Bjork, 2006),

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

which postulate that after a long period of non-use, it will be difficult to retrieve once learned material. The "New Theory of Disuse" states that after a period of non-use or in non-routine situations, the access to memories (retrieval strength) decreases even if the storage strength is high. With this in mind, studies recommend overlearning (Driskell, Willis, & Copper, 1992) or refresher interventions for tasks with long periods of non-use and for non-routine situations (Kluge, Burkolter, & Frank, 2012; Kluge & Frank, 2014).

Work experience and work performance

In an ordinary work situation, when there is no opportunity to refresh a skill, operators' work performance is influenced by their work experience (duration of employment) (Kolb, 1984; Quiñones, 2004; Tesluk & Jacobs, 1998). Work experience can be defined as the "qualitative (level of specification) and quantitative components (e.g. duration) (...) which interact and accrue over time" (Tesluk & Jacobs, 1998, p. 321). A further factor which affects work performance, irrespective of refresher interventions or work experience, is cognitive ability (Bosco & Allen, 2011; Tesluk & Jacobs, 1998).

Retentivity and skill retention in process control

Retentivity as a facet of intelligence is described as the ability to memorise information in the short- and medium term and to recognise and reproduce this information (Jäger, Süß, & Beauducel, 1997; Kersting, Althoff, & Jäger, 2008; Thurstone, 1938). Jäger (1984) defines retentivity as operative ability, which is categorised into three content abilities: Verbal thinking, numerical thinking and figural thinking. The successful memory recall and positive transfer effect of learned skills and knowledge (Baldwin & Ford, 1988; Baldwin, Ford, & Blume, 2009) depends on cognitive abilities such as retentivity (Butler, 2010; Chase & Ericsson, 1982). These appear to be generally important in controlling complex systems (Kluge, Sauer, Schüler, & Burkolter, 2009; Wittmann & Hattrup, 2004). Moreover, other person-related variables, such as self-regulation, emotional stability, and gregariousness, are also described as predictors of effective performance in process control (Xiang, Xuhong, & Bingquan, 2008). In the context of skill retention with refresher interventions, Maafi (2013) found high correlations between retentivity and performance in a simulated process control task after a longer period of non-use.

The objective of study 1 was to investigate the impact of the cognitive ability variable retentivity on training performance (skill acquisition), while study 2 investigates the impact of retentivity (Maafi, 2013) on skill retention in an ordinary process control work task (Kluge, Frank, & Miebach, 2014).

As outlined above, retentivity can be divided into verbal, numerical and figural thinking (Jäger, 1984). Verbal thinking is important, for instance, for language skills, numerical thinking for mathematical skills, and figural thinking for spatial skills. Accordingly, four retentivity measures were investigated to analyse which content ability of retentivity (Jäger, 1984) is important for skill acquisition and retention in a simulated process control task. On the basis of the available literature, the following hypotheses were developed:

In summary, it is assumed that retentivity affects skill acquisition (study 1). A group that is exposed to ordinary work experience will show less skill decay than a group without ordinary work experience (study 2). Moreover, we assume that retentivity affects skill retention (study 2), and that work experience and retentivity have an impact on skill retention (study 2).

Study 1 – Retentivity and training performance

In December 2013, the following retentivity measures were compared and evaluated with regard to their predictive validity in the context of skill acquisition in a process control task: The Selective Reminding Test (SRT), the Intelligence Structure Test 2000R (I-S-T 2000R), the Wilde Intelligence Test-2 (WIT-2) and Map Learning. The selected tests cover verbal, numerical and figural retentivity for investigating the role of retentivity in a simulated process control task.

Method

Participants
18 participants from the Engineering Department of the University of Duisburg-Essen took part in study 1. Participants were recruited by internet advertisements and flyers at the University of Duisburg-Essen (the recruitment procedure was similar for the subsequent study). All of them received course credits for their participation. They were informed about the purposes of the study and were told that they could discontinue participation at any time (in terms of informed consent).

The simulated process control task: WaTrSim
The process control task consisted of operating a Waste Water Treatment Simulation (WaTrSim; Figure 7) by applying a fixed sequence of eleven steps (Kluge & Frank, 2014). The start-up of a plant is assumed to be a non-routine task which requires skill retention (Kluge et al., 2014). In WaTrSim, the operator's task is to separate waste water into fresh water and solvent by starting up, controlling and monitoring the plant. The goal is to maximize the amount of purified water and to minimize the amount of waste water. This is achieved by controlling four main processes in WaTrSim, considering the timing of actions and following fixed sequences (Kluge et al., 2012; Kluge et al., 2014). The start-up procedure was used to measure skill retention or skill decay.

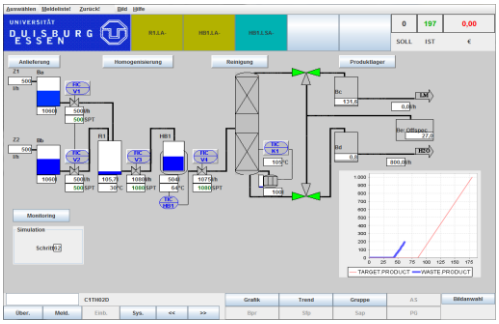


Figure 7. Interface of WaTrSim

Table 4. Sequence of start-up procedure: V1-V4 are abbreviations of valves 1-4, tanks in WaTrSim are called Ba, Bb, Bc, Bd, Be, R1 and HB1, and heating is labelled as H1 and K1 (Kluge et al., 2014)

Step #	**Temporal Transfer** (in initial training, trained start-up procedure)
Step 1	**Deactivate follow-up control** Operate controller V2 Set the target value from external to internal
Step 2	**Valve V1: Flow rate 500 l/hr** Operate controller V1 Set target value 500l/h
Step 3	**Wait until content of R1 > 200 l/hr**
Step 4	**Valve V2: Flow rate 500 l/hr** Operate controller V2 Set target value 500l/h
Step 5	**Wait until content R1 > 400 l/hr**
Step 6	**Valve V3: Flow rate 1000 l/hr** Operate controller V3 Set target value 1000 l/hr
Step 7	**Wait until content of HB1 > 100 l/hr**
Step 8	**Switch on heating H1** Operate controller HB1 set from manual to automatic operation
Step 9	**Wait until HB1 > 60°C**
Step 10	**Put column C1 into operation** Operate controller C1 set from manual to automatic operation
Step 11	**Valve V4: Flow rate 1000 l/hr** Operate controller V4 Set target value 1000 l/hr

Procedure

All participants took part in initial training (IT; Figure 8 and Table 5). The IT lasted for 120 minutes and was performed in single sessions. Participants were welcomed and introduced to WaTrSim. After completing tests assessing person-related variables and retentivity, participants explored the simulation twice. They were then given information and instructions about the start-up procedure and practised performing the target 11-step start-up procedure four times. During these first four trials, participants were allowed to use a manual which contains the eleven steps for the start-up procedure. Following this, they had to perform the start-up procedure (Table 4) four times without the manual and were told that they *were expected to produce a minimum of* **1000** *litres/hr of purified water*.

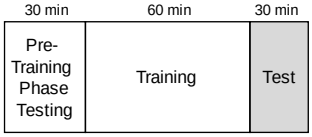


Figure 8. Initial Training (IT)

Table 5. Overview of experiment and variables of study 1

Initial Training; 120 min
• Pretraining Phase Testing: - Sociodemographic data - Retentivity tests - Previous knowledge • Initial Training: - 2x Exploration - 4x Start-up with manual • Test: - 4x Start-up without manual (*performance in final of four trials was measured*)

Measures

Predictor: Retentivity was measured using the following tests.

Selective Reminding Test (Ruff, Light, & Quayhagen, 1989): The SRT is a verbal retentivity test and consists of twelve words which had to be learned by the participants individually. After two minutes, participants had to spontaneously recall the words (without being previously aware that they would be asked to do so). If any words were missing, they had to recall these words again until they correctly recalled all twelve words on three consecutive trials or until twelve trials had been completed. After one hour, participants had to remember the words in one trial (number of words (0-12) were counted).

Intelligence Structure Test 2000R (I-S-T; Liepmann, Beauducel, Brocke, & Amthauer, 2007): The subtest "retentivity" of the I-S-T 2000R measures verbal and figural retentivity. After one minute of memorising words, the memorised words had to be matched to presented hypernyms such as "The word with an initial letter B was: a) sport, b) food, c) city, d) job or e) building" (score 0-10). After another minute of memorising, one figure of the pair was presented and the related figure had to be selected: "Please find the right answer" (score 0-13, overall score 0-23). Retentivity measured with the I-S-T 2000 R is assumed to be "low" when participants score from 0-15, "medium" for scores from 16-17, and "high" for scores from 18-23.

Wilde Intelligence Test-2 (Kersting et al., 2008): The subtest of the WIT-2 measures verbal, numerical and figural retentivity. Participants had to memorise 13 descriptions, graphics or symbols within four minutes. After a 17-minute disruption phase, they had to choose the correct solution from six alternatives in a reproduction test. The total score varied from 0-21. Retentivity measured by the WIT-2 is assumed to be low for scores from 0-12, medium for 13-14 and high for 15-21.

Map Learning (Galea & Kimura, 1993): Based on Galea and Kimura (1993), a Map Learning test measuring verbal and figural retentivity with one route, 22 objects and 20 streets on the map was imitated. The instructor showed the

participants a route, which they had to learn in under a minute. Then they had to correctly recall the route twice in succession. Mistakes were directly corrected by the instructor. After learning the route, participants were given two minutes to learn the whole map with no special instructions. They were then required to recall objects on the route, objects which were not on the route, and street names. The number of trials required to recall the route (mininum 2), the objects on the route (0-8), the objects not on the route (0-14) and the street names (0-20) were counted. A total score of recalled objects (objects on the route/not on the route and street names) was calculated (0-28).

Criterion: Performance in the start-up procedure was measured according to production outcome (purified waste water). The fourth and final trial of this series was used as the reference level of performance (production outcome) after training.

Results

The descriptive statistics are provided in Table 6.

Table 6. Descriptive statistics of predictors and criteria; M (SD), Range

Variable	*M (SD)*, Range
Sex	13 female, 5 male
Age	20.89 (2.11), 18-25
SRT (Ruff et al., 1989)	11.67 (0.59), 10-12
I-S-T 2000R (Liepmann et al., 2007)	19.33 (3.34), 9-23
WIT-2 (Kersting et al., 2008)	15.94 (1.89), 11-18
Map Learning (Galea & Kimura, 1993)	
Trials for route recall	3.28 (0.96), 2-5
Objects on the route	4.39 (1.46), 2-7
Street names	6.94 (2.58), 3-10
Total recalled objects	19.50 (3.24), 13-25
Production Outcome IT	1030.57 (127.86), 731.80-1194.59

Retentivity affects skill acquisition

*Table 7. Spearman correlation of retentivity measures as predictors and performance measures as criteria; **p<.01, *p<.05*

	1	2	3	4	5	6	7
SRT (1)	-						
I-S-T 2000R (2)	.649**	-					
WIT-2 (3)	.014	.192	-				
Map Learning							
Trials for route recall (4)	-.090	-.056	-.180	-			
Recall of objects on the route (5)	.368	.603**	.408	-.119	-		
Street names (6)	.464	.208	-.099	-.160	-.024	-	
Total recalled objects (7)	.591**	.293	.153	-.051	.324	.632**	-
Production Outcome (8)	.009	-.129	-.141	-.124	.043	.391	.503*

A Spearman correlation showed significant, medium sized correlations between total recalled objects in Map Learning and production outcome (r_s=.503, p=.033), see Table 7. A significant medium-sized significant Spearman correlation was found between production outcome and I-S-T 2000R (r_s=.506, p=.032) when the I-S-T 2000R score was divided into low, medium and high score. No significant correlations between production outcome and the other retentivity tests were found.

Discussion

Study 1 reflects direct training success and shows that retentivity measured with Map Learning and the I-S-T 2000R correlates significantly with the skill acquisition of a process control task directly after the training. In order to interpret these results, it should be added that study 1 included one measurement time (IT) only and that these two measures (I-S-T 2000R and Map Learning) address the direct recall of what was learned several minutes previously. The SRT shows no correlations with performance, which might be attributable to the verbal nature of the test, as it does not completely fit with the figural aspects of a process control task. With respect to the present results and the findings of Maafi (2013), which indicated that the I-S-T 2000R and WIT-2 are valid retentivity predictors in a process control task, in study 2, the I-S-T 2000R, Map Learning and WIT-2 (recall after 17 minutes) were used to investigate skill retention.

Study 2 – Retentivity and Skill Retention

Study 2 was conducted from March to June 2014, and investigated the impact of ordinary work experience and retentivity on skill retention in WaTrSim with four measurement times. The simulated process control task and the fixed sequence of starting up the plant was the same as described in study 1 (Table 1).

Method

Participants
38 participants took part in study 2: 18 participants in the work experience-experimental group (EG) and 20 participants in the control group (CG). The participants were recruited and instructed as described in study 1 section.

Procedure
Participants of the EG took part in initial training (IT, see above), two sessions of "ordinary work experience" (WE), and a retention assessment (RA; Figure 9), while the control group received no WE. At all measurement times the participants were tested in pairs. The IT took place as described in study 1, but was extended by a knowledge test addressing declarative and procedural knowledge at the end of the IT. The WE consisted of controlling WaTrSim twice for 30 minutes between the IT and the RA. The WE took place one week and two weeks after the IT (Kluge et al., 2014). After three weeks, the RA was conducted, lasting for approximately 20 minutes. After the participants had been welcomed, they were asked to start up the plant two consecutive times. The knowledge test which was applied at the end of the IT was also applied at the end of the RA (Table 8).

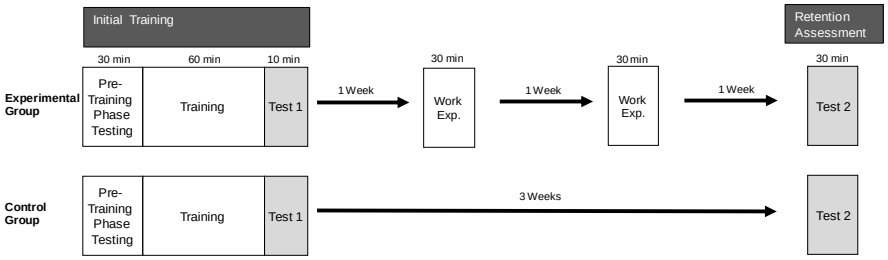


Figure 9. Procedure of study 2; the experimental group received ordinary work experience (abbr. "Work Exp.") and the control group received no work experience

Table 8. Overview of experiment parts and variables of study 2

Session Week 1 Initial Training (IT); 120 min	**Session Weeks 2 & 3** *EG only* each 30 min	**Session Week 4** Retention Assessment (RA); 30 min
• Pretraining Test: - Sociodemographic data - Retentivity - Previous knowledge • Initial Training: - 2x Explore - 4x Start-up with manual • Test 1: - 4x Start-up without manual (*performance in final of four trials was measured*) - Knowledge test	• Work Experience Task - Ordinary Work Task by controlling WaTrSim	• Test 2: - Start-up *(performance in first of two trials was measured)* - Knowledge test

Measures

Independent variable: In study 2, the EG participants took part in two simulated *work experience* (WE) sessions. The WE simulates a work day which does not including practising special skills relevant for the start-up procedure and does not contain an expected target production. The aim of WE is to continuously separate waste water into purified water and solvent. The WE consisted of the "morning scenario" and the "afternoon scenario", which have to be controlled for 30 minutes each between the IT and RA. Both scenarios took 480 seconds each. The participants were introduced to the work experience scenario with the following description "your shift starts in the morning and you take over the already running plant. The operations are manageable, but in the morning more waste water is delivered than in the afternoon. The tanker delivers 1200 litres of waste water and the valves have a flow rate of 900litres/hours". The goal of the participants was to maintain a consistent production level. They had the possibility to use the manual, which included a recommended scenario procedure (Table 9). The production outcome was measured in litres.

As a further independent variable, retentivity was measured using the *I-S-T 2000R, WIT-2* and *Map Learning* (described above).

Table 9. Example strategy for work experience scenarios "morning" and "afternoon"

Step #	Work experience scenario "morning"	Work experience scenario "afternoon"
Step 1	**Deactivate follow-up control** Operate controller V2 Set the target value from external to internal	**Deactivate follow-up control** Operate controller V2 Set the target value from external to internal
Step 2	**Valve V2: Flow rate 600 l/hr** Operate controller V2 Set target value 600l/h	**Valve V2: Flow rate 500 l/hr** Operate controller V2 Set target value 500l/h
Step 3	**Wait until content of HB1 > 100 l**	**Wait until content of HB1 > 100 l**
Step 4	**Switch on heating H1** Operate controller HB1 Set from manual to automatic operation	**Switch on heating H1** Operate controller HB1 Set from manual to automatic operation
Step 5	**Valve V3: Flow rate 720 l/hr** Operate controller V3 Set target value 720l/h	**Valve V3: Flow rate 720 l/hr** Operate controller V3 Set target value 720l/h
Step 6	**Wait until HB1 > 60°C**	**Wait until HB1 > 60°C**
Step 7	**Put column C1 into operation** Operate controller C1 Set from manual to automatic operation	**Put column C1 into operation** Operate controller C1 Set from manual to automatic operation
Step 8	**Valve V4: Flow rate 1080 l/hr** Operate controller V4 Set target value 1080l/h	**Valve V4: Flow rate 900 l/hr** Operate controller V4 Set target value 900l/h
Step 9	**Valve V1: Flow rate 600 l/hr** Operate controller V1 Set target value 600l/h	*Simulation step: 150* **Valve V1: Flow rate 500 l/hr** Operate controller V1 Set target value 500l/h
Step 10	**Valve V3: Flow rate 1200 l/hr** Operate controller V3 Set target value 1200l/h	**Valve V3: Flow rate 1080 l/hr** Operate controller V3 Set target value 1080l/h
Step 11	*Simulation step: 180* **Valve V3: Flow rate 800 l/hr** Operate controller V3 Set target value 800l/h	*Simulation step: 240* **Valve V4: Flow rate 720 l/hr** Operate controller V4 Set target value 720l/h
Step 12	*Simulation step: 320* **Valve V3: Flow rate 1080 l/hr** Operate controller V3 Set target value 1080l/h	*Simulation step: 300* **Valve V3: Flow rate 900 l/hr** Operate controller V3 Set target value 900l/h
Step 13		*Simulation step: 400* **Valve V3: Flow rate 1080 l/hr** Operate controller V3 Set target value 1080l/h

Dependent variables: The performance in the IT and RA was measured with the following variables (the fourth and final trial of the IT was used as the reference level of performance after training, and the first trial in the RA was used to assess skill retention/decay). The outcomes of the IT and RA were used for the repeated measures ANOVA and difference scores (delta) of IT and RA were used to calculate correlations and regressions:

- Production outcome, which equals the amount of purified waste water at IT, RA and Δ of IT and RA (measured in litres)
- Start-up time at IT, RA and Δ of IT and RA (time to finish the start-up procedure; max. 180 sec)
- Total number of start-up mistakes at IT, RA and Δ of IT and RA (summarised procedure and valve adjustment mistakes; 0-11)
- Procedure start-up mistakes at IT, RA and Δ of IT and RA (mistakes in steps of procedure e.g. if step 2 was taken before step 1 was executed; 0-7)
- Valve adjustment start-up mistakes at IT, RA and Δ of IT and RA (the valve flow rate was not regulated as described in the manual e.g. at 600 litres instead of 500 litres; 0-4)
- Knowledge test, which addressed declarative and procedural knowledge about WaTrSim. The test included cloze tasks, questions and diagrams about WaTrSim and background knowledge about waste water treatment (23 questions) e.g. “What are the goals in the start-up procedure in WaTrSim?”, “Which gadget is shown in the diagram?” or “Is it correct that tank R1 has to be filled with at least 100 litres so that the heating HB1 can be turned on?” (0-47)

Results

Table 7 shows the descriptive statistics and the group differences for each dependent variable. No significant differences between the groups were found (p>.05). After the experiment, the groups differed significantly in production outcome, start-up time and serious start-up mistakes in the RA (Table 7).

Table 10. Descriptive statistics of independent variables, dependent variables and control variables

Variable	Work Experience EG *M* (*SD*), Range	CG *M* (*SD*), Range	Chi² differences in prod. outcome ANOVA - group differences for each dependent variable
Control and Moderator Variables			
Sex	9 female, 9 male	10 female, 10 male	$X^2(35)=35.33$, $p=.452$
Age	25.06 (1.39), 22-28	24.65 (2.13), 21-28	$F(1,36)=0.47$, $p=.498$, $\eta^2_p=.013$
I-S-T 2000R (Liepmann et al., 2007)	19.67 (2.30), 15-23	18.20 (2.82), 12-22	$F(1,36)=3.04$, $p=.090$, $\eta^2_p=.078$
WIT-2 (Kersting et al., 2008)	12.83 (2.31), 9-17	12.60 (3.08), 6-18	$F(1,36)=0.07$, $p=.795$, $\eta^2_p=.002$
Map Learning (Galea & Kimura, 1993)			
Trials for route recall	3.33 (1.14), 2-5	3.80 (0.89), 2-5	$F(1,36)=2.00$, $p=.166$, $\eta^2_p=.053$
Objects on the route	5.22 (1.17), 3-7	4.40 (1.85), 1-8	$F(1,36)=2.62$, $p=.114$, $\eta^2_p=.068$
Street names	7.44 (3.85), 3-17	5.20 (3.02), 0-12	$F(1,36)=4.04$, $p=.052$, $\eta^2_p=.101$
Dependent Variables of IT			
Production outcome	1065.73 (194.29) 788.39-1531.69	1145.21 (103.24), 989.38-1309.61	$F(1,36)=2.55$, $p=.119$, $\eta^2_p=.066$
Start-up time	71.56 (18.54), 34-96	68.90 (10.47), 49-84	$F(1,36)=0.30$, $p=.585$, $\eta^2_p=.008$
Total start-up mistakes	1.50 (1.58), 0-4	1.15 (0.99), 0-3	$F(1,36)=0.68$, $p=.414$, $\eta^2_p=.019$
Procedure mistakes	1.11 (1.37), 0-4	0.75 (0.85), 0-2	$F(1,36)=0.98$, $p=.330$, $\eta^2_p=.026$
Valve adjustment mistakes	0.56 (0.86), 0-2	0.4 (0.88), 0-3	$F(1,36)=0.30$, $p=.585$, $\eta^2_p=.008$
Knowledge test	36.83 (3.70), 31-43	35.55 (5.00), 26-45	$F(1,37)=0.80$, $p=.378$, $\eta^2_p=.022$
Dependent Variables of RA			
Production outcome	994.39 (337.13), 189.00-1363.61	604.75 (389.59), 0.00-1066.58	**$F(1,36)=10.75$, $p=.002$, $\eta^2_p=.230$**
Start-up time	70.44 (18.98), 47-103	87.15 (35.17), 0-160	**$F(1,36)=3.21$, $p=.081$, $\eta^2_p=.082$**
Total start-up mistakes	3.78 (1.60), 1-7	4.10 (2.49), 0-11	$F(1,36)=0.23$, $p=.635$, $\eta^2_p=.006$
Procedure mistakes	2.72 (1.02), 1-5	3 (1.59), 0-7	$F(1,36)=0.40$, $p=.531$, $\eta^2_p=.011$
Valve adjustment mistakes	1.06 (1.11), 0-4	1.1 (1.48), 0-4	$F(1,36)=0.01$, $p=.918$, $\eta^2_p=.000$
Knowledge test	34.72 (4.52), 26-42	32.65 (5.35), 24-41	$F(1,37)=1.64$, $p=.208$, $\eta^2_p=.044$
Delta of IT and RA			
Production outcome	71.34 (290.98)	540.47 (382.26)	**$F(1,36)=17.80$, $p<.001$, $\eta^2_p=.331$**
Start-up time	-1.44 (17.72)	-18.5 (33.47)	$F(1,36)=3.73$, $p=.061$, $\eta^2_p=.094$
Total start-up mistakes	-2.28 (2.24)	-2.95 (2.58)	$F(1,36)=0.73$, $p=.400$, $\eta^2_p=.020$
Procedure mistakes	-1.61 (1.58)	-2.25 (1.68)	$F(1,36)=1.45$, $p=.236$, $\eta^2_p=.039$
Valve adjustment mistakes	-0.5 (1.58)	-0.7 (1.63)	$F(1,36)=0.15$, $p=.704$, $\eta^2_p=.004$
Knowledge test	2.11 (2.99)	2.9 (4.41)	$F(1,36)=0.41$, $p=.528$, $\eta^2_p=011$

Testing the hypothesis: A group that is exposed to ordinary work experience shows less skill decay than a group without ordinary work experience

In the following, repeated measures ANOVAs (for measurement time 1 and 2) with the between factor EG and CG were calculated with the dependent variables

production outcome, start-up time, total start-up mistakes, procedure start-up mistakes, valve adjustment start-up mistakes, serious start-up mistakes and knowledge test at two measurement points (results of IT are used as measurement time 1 and results of RA are used as measurement time 2). The repeated measures ANOVAs were conducted to show the skill retention or -decay between the measurement times (IT and RA) of the dependent variables.

Production outcome: A significant effect of time ($F(1,36)=30.28$, $p<.001$, $\eta^2_p=.457$), a significant effect of group ($F(1,36)=4.62$, $p=.038$, $\eta^2_p=.114$) and a significant interaction of time and group were found ($F(1,36)=17.80$, $p<.001$, $\eta^2_p=.331$; Figure 10).

Start-up time: A marginally significant effect of time ($F(1,36)=3.59$, $p=.066$, $\eta^2_p=.091$), no significant effect of group ($F(1,36)=1.43$, $p=.240$, $\eta^2_p=.038$) and a significant interaction of time and group were shown ($F(1,36)=4.58$, $p=.039$, $\eta^2_p=.113$; Figure 10).

Total start-up mistakes: A significant effect of time ($F(1,36)=43.85$, $p<.001$, $\eta^2_p=.549$) but no effect of group or interaction were found ($p>.05$).

Procedure start-up mistakes: A significant effect of time ($F(1,36)=52.95$, $p<.001$, $\eta^2_p=.595$) but no effect of group or interaction were found ($p>.05$).

Valve adjustment start-up mistakes: A significant effect of time ($F(1,36)=4.42$, $p=.043$, $\eta^2_p=.109$) but no effect of group or interaction were found ($p>.05$).

Knowledge test: A significant effect of time ($F(1,36)=16.24$, $p<.001$, $\eta^2_p=.313$) but no significant effect of group nor interaction were shown ($p>.05$).

In summary, the EG produced significantly more purified waste water and needed less start-up time than the CG. This means that the EG showed significantly less skill decay than the CG, which received no ordinary work experience, and that ordinary work experience has an impact on the performance in a process control task.

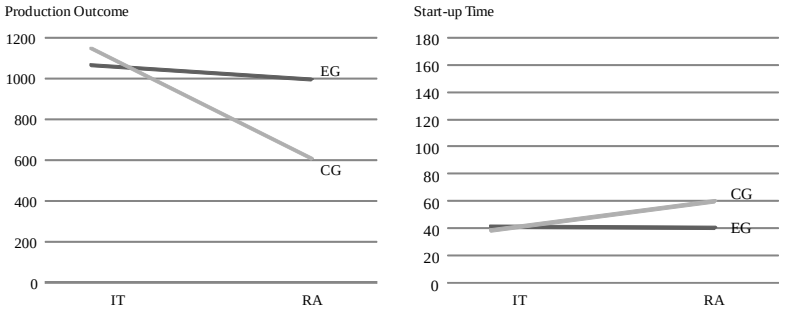


Figure 10. Production outcome (significant effect of time and interaction) and start-up time (marginally significant effect of time and significant interaction) at IT and RA of EG and CG

*Table 11. Pearson correlation between predictor retentivity measures and delta (Δ) of criteria of IT and RA (difference of IT and RA); **p<.01, *p<.05*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
I-S-T 2000R (1)	-															
WIT-2 (2)	.434**	-														
Map Learning																
Trials for route recall (3)	-.332*	-.248	-													
Recall objects on the route (4)	.281	.185	-.154	-												
Street names (5)	.213	.262	-.321*	.337*	-											
RA production outcome (6)	.174	**.392***	-.166	**.374***	.289	-										
RA start-up time (7)	.131	-.152	.218	-.240	-.088	-.431**	-									
RA total start-up mistakes (8)	-.189	-.202	.283	.054	.105	-.367*	-.089	-								
RA procedure start-up mistakes (9)	-.133	-.227	.155	-.026	.092	-.386*	-.182	.784**	-							
RA valve adjustment start-up mistakes (10)	-.161	-.085	.287	.112	.071	-.181	.046	.770**	.208	-						
RA knowledge test (11)	.222	**.477****	**-.512****	.165	.213	.383*	-.238	-.251	-.212	-.177	-					
Δ production outcome (12)	-.215	**-.347***	.201	**-.327***	-.207	-.928**	.367*	.471**	.488**	.241	-.335*	-				
Δ start-up time (13)	-.171	.206	-.088	.212	.017	.258	-.819**	.051	.039	.041	.169	-.291	-			
Δ total start-up mistakes (14)	.195	.116	-.261	-.133	-.005	.380*	.044	-.844**	-.585**	-.730**	.200	-.401*	-.166	-		
Δ procedure start-up mistakes (15)	.137	.161	-.242	-.109	.016	.371*	.050	-.664**	-.734**	-.292	.124	-.417**	-.067	.783**	-	
Δ valve adjustment start-up mistakes (16)	.177	.059	-.177	-.052	.005	.236	-.006	-.641**	-.166	-.840**	.205	-.242	-.172	.721**	.158	-
Δ knowledge test (17)	-.126	-.138	.246	.086	-.119	.056	-.003	-.160	-.002	-.251	-.527**	-.026	.001	.179	.017	.240

Note: *Deltas can result in negative correlations*

Testing the hypothesis: Retentivity affects skill retention

A Pearson correlation showed significant, moderate correlations (p<.05) between WIT-2 and production outcome at RA, WIT-2 and knowledge test at RA, Map Learning and production outcome at RA, and Map Learning and knowledge test at RA. Additionally, moderate correlations between WIT-2 and the delta of production outcome (difference of IT and RA) and between Map Learning and the delta of production outcome were found, as shown in Table 11.

In summary, the findings show that performance in the RA correlates significantly with retentivity, and that the IT-RA difference score (Δ) of performance correlates significantly with retentivity.

Testing the hypothesis: Work experience and retentivity have an impact on skill retention

A regression was conducted to investigate the impact of both independent variables (ordinary work task and retentivity) on skill retention. The model with predictors group and WIT-2 on criterion delta of production outcome explained a 43.5% of the variance (F(2,35)=13.48, p<.001; Table 12). A regression with the predictors group and I-S-T 2000R on criterion delta of production outcome explained 33.4% of the variance (F(2,35)=8.78, p=.001; Table 12). Furthermore, a regression with the predictors group and Map Learning (objects on the route) on criterion delta of production outcome resulted in a significant model, which explained 36.4% of the variance (F(2,35)=10.03, p<.001; Table 12). A regression with the predictors group

and Map Learning (trials for route recall) explained 33.6% of the variance ($F(2,35)=8.85$, $p=.001$; Table 12).

Finally, a regression with the predictors group and WIT-2/group and I-S-T 2000R/group and Map Learning on criterion delta of knowledge test showed no significant model (p>.05).

The results indicate that the significant model with the predictors work experience and WIT-2 explains the greatest amount of variance and that both predictors have a significant impact on the criterion variable.

Table 12. Regression with criterion variables production outcome and knowledge test

Criterion variable: Delta of production outcome IT and RA					
Predictor	*B*	*SE(B)*	*ß*	*T*	*p*
Group	**457.650**	**103.709**	**.561**	**4.413**	**<.001**
WIT-2	**-49.179**	**19.361**	**-.323**	**-2.540**	**.016**
Criterion variable: Delta of production outcome IT and RA					
Group	455.63	117.14	.559	3.89	<.001
I-S-T 2000R	-9.20	22.29	-.059	-0.41	.682
Criterion variable: Delta of production outcome IT and RA					
Group	428.833	113.830	.526	3.767	.001
Map Learning: Objects on the route	-49.004	36.076	-.190	-1.358	.183
Criterion variable: Delta of production outcome IT and RA					
Group	455.52	115.41	.559	3.95	<.001
Map Learning: Trials	29.15	56.69	.073	0.51	.610

Discussion

The objective of the second study was to investigate skill retention in an ordinary work task and the impact of the cognitive ability variable retentivity on performance in a process control task.

Study 2 showed that the EG outperformed the CG in the production of purified waste water and starting up the plant. This suggests that operating the plant and having work experience is more supportive than having no interaction with the system (Kluge et al., 2014). In addition, the study shows that retentivity measured by WIT-2 and Map Learning correlates with skill retention in process control tasks, with medium effect sizes. The regressions with production outcome as criterion variable showed significant results for all predictors, but the model with group and WIT-2 as predictors was the only model in which both variables had a significant impact on the criterion. This suggests that work experience and the retentivity measure WIT-2 can be used as retentivity measures in simulated process control tasks, which is in accordance with Maafi (2013).

General discussion

In general, the results suggest that work experience positively affects skill retention (Kluge et al., 2014) and that retentivity as an individual difference can predict work

performance (Tesluk & Jacobs, 1998) and skill retention. In addition, the results show that the simulated process control task addresses verbal, numerical and figural retentivity (Jäger, 1984; Jäger et al., 1997), which can be measured with WIT-2: "verbal retentivity" by remembering the labels of a tank, "numerical" by remembering the rate of flow of a valve, and "figural" by remembering the symbols for tanks, valves or column and the arrangement of the symbols. The findings also demonstrate that in terms of skill retention with two measurement times, retentivity should be measured using a test comprising two measurement times.

Limitations and implications

The present studies were implemented in a micro-world setting and using a student sample. It is possible that the study was limited due to the special-purpose experimental setting (Stone-Romero, 2011). Additionally, in order to investigate participants who were as similar as possible to the operators to whom we wish to generalise the findings, engineering students were recruited for the study. Finally, it is virtually impossible to investigate these purposes in a real process control setting, and in particular to recruit 40 almost identical operators with the same level of training and experience, and, in order to conduct a controlled and valid experiment, to bring real operators to the lab four times.

The present findings and previous studies (Kluge et al., 2014; Maafi, 2013) show that future research on retentivity and skill retention would be worthwhile. It would be interesting to investigate retentivity in the context of work experience in comparison to refresher interventions, and in the refresher context only (Kluge et al., 2012; Kluge & Frank, 2014). In future experiments, it would be recommendable to investigate general mental ability and its impact on retentivity and to recruit a larger sample size.

Practical implications

The findings indicate that the cognitive ability variable retentivity is a valid predictor of skill retention. In addition, it suggests that the WIT-2 provides a good possibility to measure retentivity in process control tasks in only 20 minutes. Therefore, it can be recommended as one instrument for the selection of personnel for process control.

Acknowledgements

The studies were carried out with the help of Julia Miebach (study 1) and Marcel Reefmann (Study 2). We thank both of them for their assistance.

References

Bainbridge, L. (1983). Ironies of automation. Increasing levels of automation can increase, rather than decrease, the problems of supporting the human operator. *Automatica*, *19*, 775-779.

Baldwin, T.T., & Ford, J.K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, *41*, 63-105.

Baldwin, T.T., Ford, J.K., & Blume, B.D. (2009). Transfer of training 1988–2008: An updated review and agenda for future research. In G. P. Hodgkinson, & J. K. Ford (Eds.), *International review of industrial and organizational psychology* (pp. 41-70). Chichester: John Wiley & Sons, Ltd.

Bjork, R.A., & Bjork, E.L. (2006). Optimizing treatment and instruction: Implications of a new theory of disuse. In L.G. Nilsson, and N. Ohta (Eds.), *Memory and society. Psychological perspectives* (pp. 109-134). Hove: Psychology Press.

Bjork, R.A., & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, and R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 35-67). Hillsdale: Erlbaum.

Bosco, F.A., & Allen, D.G. (2011). Executive attention as predictor of employee performance. *Academy of Management Proceedings, 11*, 1-6.

Bourne, L., & Healy, A. (2012). Introduction: Training and its cognitive underpinnings. In A.F. Healy, and L.E. Bourne (Eds.), *Training cognition. optimizing efficiency, durability, and generalizability* (pp. 1-12). New York: Psychology Press.

Butler, A.C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*, 1118-1133.

Chase, W.G., & Ericsson, K.A. (1982). Skill and working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation* (pp. 1-58). London: Academic Press.

Driskell, J.E., Willis, R.P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology, 77*, 615-622.

Galea, L.A., & Kimura, D. (1993). Sex differences in route-learning. *Personality and Individual Differences, 14*, 53-65.

Jäger, A.O. (1984). Intelligenzstrukturforschung: Konkurrierende modelle, neue entwicklungen, perspektiven. *Psychologische Rundschau, 34*, 21-35.

Jäger, A.O., Süß, H., & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test, Form 4*. Göttingen: Hogrefe.

Kaber, D.B., Omal, E., & Endsley, M. (1999). Level of automation effects on telerobot performance and human operator situation awareness and subjective workload. *Human Factors and Ergonomics in Manufacturing, 10*, 409-430.

Kersting, M., Althoff, K., & Jäger, A. (2008). *Wilde-Intelligenz-Test 2 (WIT-2) (Manual).* Göttingen: Hogrefe.

Kluge, A., Burkolter, D., & Frank, B. (2012). "Being prepared for the infrequent": A comparative study of two refresher training approaches and their effects on temporal and adaptive transfer in a process control task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56*, 2437-2441.

Kluge, A. (2014). *The acquisition of knowledge and skills for taskwork and teamwork to control complex technical systems*. Heidelberg: Springer.

Kluge, A., & Frank, B. (2014). Counteracting skill decay: Four refresher interventions and their effect on skill and knowledge retention in a simulated process control task. *Ergonomics, 57*, 175-190.

Kluge, A., Frank, B., & Miebach, J. (2014). Measuring skill decay in process control-results from four experiments with a simulated process control task. In D. de Waard, K. Brookhuis, R. Wiczorek, F. di Nocera, R. Brouwer, P. Barham, C. Weikert, A. Kluge, W. Gerbino, & A. Toffetti (Eds.) (pp. 79-93). *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2013 Annual Conference*. Retrieved from hfes-europe.org

Kluge, A., Sauer, J., Schüler, K., & Burkolter, D. (2009). Designing training for process control simulators: A review of empirical findings and current practices. *Theoretical Issues in Ergonomics Science, 10*, 489-509.

Kolb, D.A. (1984). *Experiential learning: Experience as the source of learning and development.* Englewood Cliffs, NJ: Prentice-Hall.

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). *Intelligenz-Struktur-Test 2000R: IST 2000 R* (Manual) (2nd ed.). Göttingen: Hogrefe.

Maafi, S. (2013). *Trainieren, Merken, Abrufen! Der Einfluss personenspezifischer Variablen auf die Trainings- und Transfereffekte in der Prozesskontrolle*. [Train, remember, recall! the impact of person-related variables on training and transfer effects in process control]. (Unpublished Master thesis). Chair of Business and Organisational Psychology, University of Duisburg-Essen, Duisburg, Germany.

Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 30*, 286-297.

Quiñones, M.A. (2004). Work experience: A review and research agenda. In C.L. Cooper, & I.T. Robertson (Eds.), *International review of industrial and organizational psychology 2004* (pp. 119-138). Hoboken: John Wiley & Sons, Ltd.

Ruff, R.M., Light, R.H., & Quayhagen, M. (1989). Selective reminding tests: A normative study of verbal learning in adults. *Journal of Clinical and Experimental Neuropsychology, 11*, 539-550.

Stone-Romero, E.F. (2011). Research strategies in industrial and organizational psychology: Nonexperimental, quasi-experimental, and randomized experimental research in special purpose and nonspecial purpose settings. In S. Zedeck (Ed.), *APA handbook of industrial and organizational psychology. volume 1, building and developing the organization* (pp. 37-72). Washington: American Psychological Association.

Tesluk, P.E., & Jacobs, R.R. (1998). Toward an integrated model of work experience. *Personnel Psychology, 51*, 321-355.

Thurstone, L.L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

Wickens, C.D., & McCarley, J.S. (2008). *Applied attention theory*. Boca Raton: CRC Press.

Wittmann, W.W., & Hattrup, K. (2004). The relationship between performance in dynamic systems and intelligence. *Systems Research and Behavioral Science, 21*, 393-409.

Xiang, F., Xuhong, H., & Bingquan, Z. (2008). Research of psychological characteristics and performance relativity of operators. *Reliability Engineering & System Safety, 93*, 1244-1249.

Implementing dynamic changes in automation support using ocular-based metrics of mental workload: a laboratory study

Serena Proietti Colonna[1], *Claudio Capobianco*[2], *Simon Mastrangelo*[2], *Francesco Di Nocera*[1]
[1]*Sapienza University of Rome*
[2]*Ergoproject s.r.l., Rome*
Italy

Abstract

Adaptive Automation has been often invoked as a remedy to indiscriminate introduction of automation support. However, this form of automation is difficult to implement without a sensitive and reliable index of the Operator Functional State. In a series of studies we have showed the usefulness of the distribution of eye fixations as an index of mental workload to be used as a trigger of automation. Particularly, the distribution pattern was found to be sensitive to taskload variations and types, thus making it very appealing for designing adaptive systems. This approach seems to be valid and reliable, but a necessary step in this research program would be testing the effectiveness of automation driven by fixation distribution and its capability in reducing the workload. The present study is a first attempt to carry out such validation.

Introduction

In many work domains the introduction of automation can improve complex systems performance and reduce overall costs by limiting the intervention of human operators. This can be accomplished in several ways: for example, through the assignment of routine tasks to computer systems in order to relieve the operator from performing them, as well as by implementing automatic monitoring of a process in order to improve safety (Rouse, 1981). Automation could also be implemented for removing the operators from dangerous work environments and for operating in environments that are inaccessible to the humans (Sheridan, 1992).

However, a major challenge in automation design is function allocation, that is "what needs to be automated" and "to what extent" in order to optimise performance (Inagaki, 2003). Several models have been devised in order to answer those questions and for supporting automation design. Some accounts represent all-purpose taxonomies initially developed in specific research domains (e.g. Sheridan and Verplank, 1978), whereas others attempted to address the issue of function allocation in terms of its relation with human information processing (e.g. Parasuraman, Sheridan & Wickens, 2000). There is, however, a third question that

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

should be answered in order to properly allocate functions, which is "when to automate". Indeed, albeit technology is aimed at reducing mental workload and errors, human interaction with automated systems also result in paradoxical side effects: automating a task often lead to loss of situation awareness and to higher mental workload when the operator is asked to intervene into the ongoing operation, particularly if the operator has been confined to monitoring functions for prolonged periods. With that in mind, it would be desirable having more flexible forms of automation in which function allocation dynamically varies during system operation and it is matched to what has been recently defined as "operator functional state" (Hockey, Gaillard, & Burov, 2003). That would facilitate a positive trade-off between the benefits and costs of automation itself (Parasuraman and Wickens, 2008). This form of automation is usually called "adaptive" and represents a closed-loop system in which the state of the operator is constantly monitored and support is provided only when it is needed.

The quest of a trigger

Adaptive automation is difficult to implement without a sensitive and reliable index of mental workload. The choice of the index to use for triggering the system when the functional state of the operator significantly deviates from optimal levels is thus one of the most important issues both for research purposes and for effective implementation of dynamic function allocation. Spontaneous psychophysiological activity showing sensitivity to variations in mental workload (e.g. cardiovascular, cerebral and ocular activity) is commonly considered the best choice, because it provides the opportunity for steady monitoring (and control). Many efforts have been devoted by several research groups for finding viable methodologies. Only to name a few, indices of "engagement" obtained from continuous EEG (Pope, Bogart, & Bartolome, 1995) or neural networks integrating data from multiple psychophysiological measures (Wilson, Lambert, & Russell, 2000) have been tested as potential triggers for adaptive systems. However, there is still no agreement in the literature about which indicator to use.

Among the many valuable accounts in the literature, our research group has proposed the use of the distribution of eye fixations as an index of mental workload and a potential trigger for adaptive systems. Particularly, a statistical indicator of spatial dispersion (the Nearest Neighbour Index) has been repeatedly found to vary with taskload (Camilli et al., 2007; 2008; Di Nocera & Bolia, 2007; Di Nocera et al. 2006; 2007; 2014). The index is based on the ratio between minimum distances observed in the distribution of eye fixations and the minimum distance expected by chance. Fixations spreading appear to be associated to mental workload when taskload depends on the temporal demand, whereas fixations clustering would be associated to mental workload when taskload depends on the visio-spatial demand. This index seems to be valid and reliable, but a necessary step in this research program would be testing the effectiveness of NNI-driven automation support in reducing the workload. Particularly, an adaptive system based on NNI should: 1) activate when the index deviates from (a previously computed) baseline; 2) produce a corresponding change in the fixation distribution (back to baseline limits)

indicating a mitigation of workload; 3) deactivate when that state has been reached. The present study is a first attempt to carry out such validation.

Study

Participants. Nineteen individuals (9 females, mean age = 26.52 st. dev. =2.65) volunteered in this study. All participants were right-handed and had normal or corrected to normal vision.

Experimental setup. The X2-30 wide eye tracking system (Tobii, Sweden) was used for recording ocular activity and custom Matlab code has been developed for running this experiment. The Tetris game, a commonly known tile-matching puzzle videogame successfully endorsed in a variety of studies (e.g., Trimmel & Huber, 1998), was used as experimental task. The gaming platform was based on "matlabtetris" by Matt Figg[2]. The entire experimental package was developed using Matlab® 2013a along with Tobii Analytics SDK v. 3.0 and was composed by three modules: the Tetris game, the NNI suite and the NNI monitor. The layout and the graphics of the game were kept as minimalistic as possible in order to reduce spurious saccades. Ocular data sampling frequency was set at the maximum available rate (40Hz). The NNI suite was created after the ASTEF package code (Camilli et al., 2008), performed all the tasks related to spatial statistics and computed them in real-time in order to trigger the automation support. This module can compute NNI based on convex-hull or smallest-rectangle areas, with or without the Donnelly adjustment. The suite can also analyse data and generate time series to be used in successive statistical analyses. The NNI monitor (available to the experimenter for visual inspection during the recording) plots the ongoing NNI value.

Procedure. Participants were seated in front of a 17" display, at a distance of approximately 60 cm. The room was dimly illuminated only by the display. After calibration of the eye-tracking system, they underwent a practice run of the Tetris game at the same velocity of the real game for avoiding context effects in the subjective assessment (see Colle & Reid, 1998).

The version of the game used in this study acted as a common version of the Tetris game with the exception that in this case the game restarted from a blank screen (starting condition) each time the stack of Tetriminos (game pieces) reached the top of the playing area and no new Tetriminos were able to enter. This condition commonly denotes the end of the game, whereas in this very situation the game needed to go on until the end of the entire experimental session (10 minutes each session). The game was therefore immediately restarted when the Tetriminos (reached the top and the restart was scored as a loss: a performance indicator to be used as dependent variable (# of restarts).

[2] http://www.mathworks.com/matlabcentral/fileexchange/34513-matlabtetris

Automation support was implemented as a projection of the falling Tetrimonos ("ghost block") over the pieces lying at the very bottom of the game area. This is known to facilitate the proper positioning by providing a time gain to the player. This manipulation has been already used in one previous study (Di Nocera et al., 2006).

During the first 5 minutes of gaming the NNI baseline for each subject was computed in real-time and any NNI value greater than ±1 standard deviation was marked as "out of range". Data collected within this "calibration" epoch was not included in the analyses.

Three automation conditions were implemented: manual control (no automation support), self-paced automation support (subjects could activate/deactivate the ghost block at their ease), and adaptive automation (the ghost block appeared when NNI deviations from baseline occurred and disappeared right after the NNI values returned within baseline limits). Each condition lasted 15 minutes and the presentation order was balanced across participants.

Data analysis and results

Given the scope of this study, the dependent variable to employ should represent the effectiveness of automation support in producing a return to NNI baseline values after deviation. With that in mind, a composite variable (proportion of "inwards" after deviation) has been computed by dividing the number of consecutive minutes within the ±1 standard deviation interval by the number of total minutes within the ±1 standard deviation interval. This measure would represent the effectiveness of the automation support in keeping the individual within acceptable workload levels for a prolonged period. The variable has been computed for all conditions (manual, self-paced, adaptive), thus we should expect a lack of significant differences between conditions if the return to baseline is random and/or "physiological". Two ANOVA mixed designs were carried out using the proportion of "inwards" (system effectiveness) and the number of game restart (individual performance) as dependent variables. Condition (Manual vs. Self-paced vs. Adaptive) and Gender (Males vs. Females) were used as factors. The latter was included in order to control for differences between males and females in computer gaming performance (see American Association of University Women, 1998). Results showed a significant interaction Gender by Condition for the proportion of inwards ($F_{2,38}$=3.10, p=.056). Duncan post-hoc testing showed that the interaction was due to males showing greater proportion of inwards after deviation with the adaptive automation than with manual control and self-paced automation ($p<.05$; figure 1). Main effects of Gender ($F_{1,19}$=6.90, $p<.05$) and a tendency towards statistical significance for Condition (p=.08) were found for the number of restart. Females gaming performance was significantly worse than males' and Duncan post-doc testing showed that gaming performance with self-paced automation was worse than that in manual control and adaptive automation.

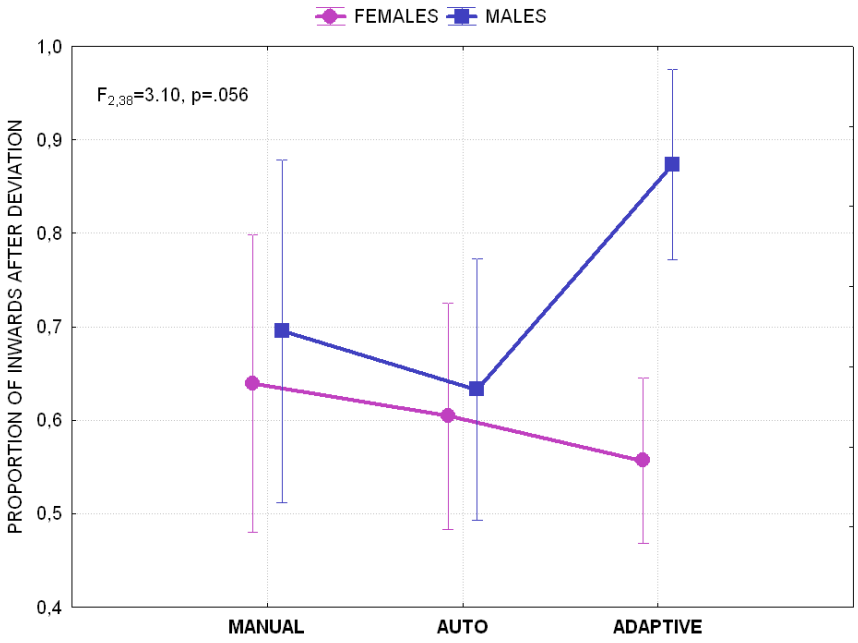


Figure 1. Proportion of inwards after deviation by condition and gender.

Discussion and Conclusions

An adaptive system should provide support for mitigating mental workload when the operator functional state is compromised and it should deactivate when the operator is back into "normal" functioning. In this study we have devised an experimental adaptive system based on the distribution of eye fixations. It was a rather basic system aimed at a laboratory investigation. The system was able to activate the automation support when the ocular index deviated from (a previously computed) baseline and to deactivate it when the index returned to baseline values. Changes in the index values obtained in the adaptive automation condition were compared to those occurring in the same task under manual control and self-paced automation support. Results showed a beneficial effect of the ocular-driven adaptive automation, but limited to male participants. Analyses carried out on gaming performance showed that females performed significantly worse than males in this task, thus suggesting that the gender difference found for the automation support should probably be considered a floor effect. Differential ability of males and females with visuo-spatial gaming and computing in general is well known (American Association of University Women, 1998) and in this case has been exacerbated by the absence of a proper training with the game prior to experimentation.

Interestingly, we found a detrimental effect of self-paced automation on performance. Apparently, performance in the adaptive automation condition

matches that obtained with manual control, although it was characterized by better workload management (as indicated by results on the proportion of inwards, even if limited to males). Self-paced automation appears to "get into the way", neither producing a mitigation of workload nor improving performance.

This was a first attempt in testing the potential of the NNI as a real-time trigger for automation. Moreover, these findings and the potential application of the technique are limited to those settings in which an operator seats in front of a display (e.g. Air Traffic Control). Results are far from being conclusive, but yet encouraging. One of the major flaws of the present study was lack of training with the task that probably affected female participants most. A replication of this study with a trained sample showing homogeneous performance levels is needed to disentangle the effect found.

Acknowledgements

This study was funded by the European Commission through the FP7 project "CyClaDes - Crew-centred Design and Operations of ships and ship systems".

References

American Association of University Women (1998). *Separated by sex: A critical look at single-sex education for girls.* Washington, DC: American Association for University Women Educational Foundation.

Camilli, M., Terenzi, M., & Di Nocera, F. (2007). Concurrent validity of an ocular measure of mental workload. In D. de Waard, B. Hockey, P. Nickel, and K. Brookhuis (Eds.), *Human Factors Issues in Complex System Performance* (pp. 117-129). Maastricht, The Netherlands: Shaker Publishing.

Camilli, M., Nacchia, R., Terenzi, M., & Di Nocera, F. (2008). ASTEF: A simple tool for examining fixations. *Behavior research methods*, *40*, 373-382.

Camilli, M., Terenzi, M., & Di Nocera, F. (2008). Effects of Temporal and Spatial Demands on the Distribution of Eye Fixations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52, 1248-1251.

Colle, H.A., & Reid, G.B. (1998). Context effects in subjective mental workload ratings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(4), 591-600.

Di Nocera, F., Terenzi, M., & Camilli, M. (2006). Another look at scanpath: distance to nearest neighbour as a measure of mental workload. In D. de Waard, K.A. Brookhuis, and A. Toffetti (Eds.), *Developments in Human Factors in Transportation, Design, and Evaluation* (pp. 295-303). Maastricht, the Netherlands: Shaker Publishing.

Di Nocera, F., & Bolia, R.S. (2007). PERT networks as a method for analyzing the visual scanning strategies of aircraft pilots. In *Proceedings of the 14th International Symposium on Aviation Psychology* (pp. 165-169).

Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: pilot's scanning strategies and the real-time assessment of mental workload. *Journal of Cognitive Engineering and Decision Making*, 1, 271-285.

Di Nocera, F., Proietti Colonna, S., Dessì, F., Capobianco, C., Mastrangelo, S., Steinhage, A. (2014). Keep Calm and Don't Move A Muscle: Motor

restlessness as an indicator of mental workload. In D. de Waard, K. Brookhuis, R. Wiczorek, F. Di Nocera, P. Barham, C. Weikert, A. Kluge, W. Gerbino, and A. Toffetti (Eds.), *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2013 Annual Conference* (pp. 183-191). Available as open source download. ISSN 2333-4959 (online).

Hockey, G.R.J., Gaillard, A.W.K., & Burov, O. (2013), Operator Functional State: The Assessment and Prediction of Human Performance Degradation in Complex Tasks. Amsterdam: IOS Press.

Inagaki, T. (2003). Adaptive automation: sharing and trading of control. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design* (pp. 147–169). London: Lawrence Erlbaum Associates.

Parasuraman, R., Sheridan T.B., Wickens, C.D. (2000). A Model for types and levels of human interaction with automation IEEE *Transactions on Systems, Man, and Cybernetics, Part A. Systems and Humans, 30 (3)*, 286–297.

Parasuraman R., Wickens C.D., (2008). Humans: still vital after all these years of automation. *Human Factors*, *50*, 511–520.

Pope, A.T., Bogart, E.H., & Bartolome, D.S. (1995). Biocybernetic system evaluates indices of operator engagement in automated task. *Biological psychology*, *40*, 187-195.

Rouse, W.B. (1981). Human-computer interaction in the control of dynamic systems. *Computing Surveys*, 13, 71-99.

Sheridan, T.B., & Verplank, W.L. (1978). *Human and computer control of undersea teleoperators*. Massachuttes Instistute Of Technology. Cambrige: Man-Machine Systems Lab.

Sheridan T.B. (1992). Telerobotics, automation, and human supervisory control. Cambridge: MIT Press.

Trimmel, M., & Huber, R. (1998). After-effects of human-computer interaction indicated by P300 of the event-related brain potential. *Ergonomics*, *41*, 649-655.

Wilson, G.F., Lambert, J.D., & Russell, C.A. (2000). Performance enhancement with real-time physiologically controlled adaptive aiding. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 61-64.

Event expectancy and inattentional blindness in advanced helmet-mounted display symbology

Patrizia Knabl, Sven Schmerwitz, & Johannes Ernst
German Aerospace Centre (DLR), Institute of Flight Guidance
Germany

Abstract

Helmet-mounted displays (HMD) have the potential to significantly increase helicopter flight safety by superimposing synthetic information in the forward field of view. Particularly during poor visibility, decreasing workload and enhancing situation awareness are two key factors. However, previous findings mainly within the scope of head-up displays in fixed-wing aviation have shown that superimposed displays also pose a risk of impairing the detection of unexpected events. The present paper will investigate this topic in the context of new HMD symbology concepts for rotary-wing aircraft. The designs were tested in a simulator study with 18 civil and military pilots. Primarily, attention distribution in terms of concurrent task performance was investigated. In addition, two unexpected events occurred, a warning on the display and a traffic incursion in the outside scene. Results revealed a later response to the warning on the HMD, if it was presented truly unexpected and in poor visibility. Moreover, a trend towards an HMD detection cost for the traffic incursion was observed.

Introduction

In recent years helmet-mounted displays became increasingly important for rotary-wing aircraft. They provide pilots with relevant flight information by presenting symbology in the forward field-of-view. Therefore head-down time and scanning costs between instruments and outside environment can be reduced. As a result, divided attention tasks are facilitated by enabling a parallel monitoring of the two domains. This advantage is essential, since maintaining constant visual contact with the environment is time-critical especially in low altitude flight and poor visibility conditions. Nevertheless, an appropriate symbology design is crucial to enhance situation awareness and reduce workload and spatial disorientation. Moreover, it has been found that event detection performance with superimposed displays is largely dependent on the expectancy of the events. Expected events are usually classified as those who are naturally expected during flight, occur frequently or have specifically been briefed. In contrast, unexpected events refer to those who occur truly surprising, rarely, are usually not anticipated or briefed. Findings in literature indicate that head-up displays usually facilitate the detection of expected events in the environment or on the display. However, costs have often been observed in the

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

detection of unexpected events, especially if they occur in the outside scene and are not very salient (Fadden, Ververs, & Wickens, 1998; Wickens & Long, 1994, 1995). For instance it was found that runway incursions during approach are detected later with a head-up display, a result that was also observed with HMDs by Lorenz, Többen, and Schmerwitz (2005). The finding can be attributed to both a cost of clutter, as well as attentional tunneling. The former hinders event detection by presenting too much information in the forward field of view and by obscuring objects with the symbology. The latter refers to an inadequately long allocation of attention to the symbology that leads to neglecting relevant events on other channels as well as failing to perform other tasks (Wickens & Alexander, 2009). These findings can further be related to the concept of inattentional blindness (Mack & Rock, 1998) since it is described as a "failure to see highly visible objects we may be looking at directly when our attention is elsewhere" (Mack, 2003, p. 180). As a result, simply superimposing symbology to assume parallel processing of information within a specific area or spotlight of attention, as adapted from space-based attention theories (Eriksen & Eriksen, 1974) has proven not to be efficient. The topic is discussed elaborately in the context of head-up displays in Wickens and Long (1995). It has rather become evident that selective attention is in fact driven by bottom-up processes such as saliency and effort, as well as top-down processes such as expectancy and value, constituting the basic components of the SEEV-model (Wickens, Helleberg, Goh, Xu, & Horrey, 2001; Wickens & McCarley, 2007). In general, Wickens and Alexander (2009) point out that the unexpected event detection cost should not lead to the overall conclusion to classify superimposed displays as being generally problematic. Moreover using conformal symbology was found to mitigate this problem. It refers to symbology that is somewhat linked with the far domain by being spatially aligned with actual or virtual objects in the environment, such as conformal horizon lines, runways or obstacles. Nevertheless it has to be noted that the previous literature very strongly focuses on head-up displays in the fixed-wing domain, or monocular HMDs with a rather small field of view. The present paper, however, investigates the use of conformal symbology in modern, binocular HMDs and focuses on low altitude and poor visibility helicopter operations. Therefore new symbology concepts for en route and landing assistance featuring conformal obstacle and route presentations were tested in a real-time simulation. Test subjects were instructed to monitor for attitude changes on the display and perform a search and identification task in the outside scene. Furthermore two unexpected events were presented. The paper subsequently focuses on the unexpected event detection results. Findings regarding the main task performance are described in Knabl, Schmerwitz, Doehler, Peinecke, and Vollrath (2014).

Method

Participants

Eighteen pilots with an average age of 45 years (SD = 7) participated in the study. Nine were military pilots from the German Armed Forces and nine were civil pilots from the German Federal Police Force and the German (DRF) and Swiss (REGA) air rescue providers. Their average flight experience was 4401 (SD = 3867) flight

hours in total and 1167 (SD = 758) on the presently operated helicopter type. 17 pilots were instrument rated. Six had experience with an HMD in the simulator, four had additional experience in real flight, and eight had never flown with an HMD before. Their HMD experience averaged 178 (SD = 366) flight hours. The aircraft types most frequently operated were the EC 135, NH 90 and AS 332.

Apparatus

Helmet-mounted display
The HMD used was a JEDEYE™ helmet system by Elbit Systems Ltd. (figure 1). It features a wide field of view (80° x 40°) and a very high resolution (2 x 1920 x 1200 px). Table 1 provides the most significant technical specifications.

Table 1. Technical specifications of the JEDEYE™ helmet system.

Resolution	2x1920x1200 pixel @ 60 Hz
Field-of-view	binocular, 2x80°x40°, stereo capable
Head tracker	magnetic, 400 Hz, precision 0.25°
Weight	approx. 2.3 kg incl. helmet
Interface	RS-170, SDI, DVI-D, HDMI
Colour space	monochrome green

Simulator
The fixed-base simulator GECO (generic cockpit simulator, figure 1) provided a collimated projection with a resolution of 3 x 2560 x 1440 pixel spreading 180° x 40°. The cockpit shell was a model of an Airbus A320, but furnished with the HMI layout of an A350. In order to allow helicopter experiments the simulator was equipped with a cyclic and collective on the right seat. Both inputs allowed active feedback. The regular yaw control was modified to have low resistance and no resilience. As flight model an EC-135 was used with the software simulation tool X-Plane10. The realism of the model was rather low but allowed easy handling for the test subjects. The cockpit shell did not provide enough forward slant view due to the high glare shield. Therefore the simulated horizon was tilted 5° upwards. None of the participants commented on having been irritated by this.

Experimental design

Each pilot conducted twelve scenarios, six with the use of the HMD and the same six with the head-down baseline condition. The visual condition as well as the display and scenario order was permuted. Half of the participants completed all scenarios with the HMD followed by the baseline scenarios and vice versa. Within

each display condition three scenarios started with average visibility changing to poor half way through the run and vice versa.

Figure 1. Generic Cockpit Simulator (GECO) and JEDEYE™ helmet system.

Procedure

The trials took place from March till May 2014 at the DLR Institute of Flight Guidance. Each pilot spent a full eight hour day at the institute. The morning session consisted of an introduction, briefing, familiarization and training, the afternoon session of testing and de-briefing. Within the training phase the participants were given time to become accustomed to the aircraft and HMD symbology. Moreover the primary tasks were trained. All test subjects signed a letter of consent and filled out a biographical questionnaire, containing questions about age, flight experience, usage of HMD, as well as experience with brownout and spatial disorientation. The actual test phase was split into two blocks, each taking approximately 80 minutes to complete, and separated by a 15 minute break. A block consisted of six scenarios with one display set. Pilots wore the HMD with the visor folded down also in the baseline scenarios to ensure equal brightness and contrast. The de-briefing collected various subjective aspects using tailor-made questionnaires with regard to helmet use and symbology design.

Symbology

Both display types (head-down and head-up) presented almost identical situation and navigation information. The head-down variant was designed according to the fielded instrumentation of DLR's EC135 helicopter ACT/FHS and was split into two screens, the primary flight display (PFD) (figure 2, top left) and the navigation display (ND) (figure 2, bottom left). In the simulator they were located directly in front of the pilot and shared a 15.4'' TFT with 1440x900 pixels. The PFD primarily delivered information on speeds, heading, heights and attitude/horizon. To maintain high transparency within the helmet display (low clutter) the representation of information was greatly simplified (figure 2, top right). Additionally the head-up symbology was presented in monochrome green whereas the head-down symbology was presented in colour. The type of "glass cockpit"-PFD was well known to almost

all pilots. The significant difference of the HMD symbol set was the combined presentation of PFD and ND into an egocentric perspective display (figure 2, bottom right). The predetermined route was visualized by route points consisting of virtual conformal, terrain-based arrows, and waypoints presented as poles. Furthermore, conformal obstacles (power lines, windmills and towers) were depicted, whereas the head-down symbology did not feature obstacle highlighting. Finally, the ND delivered route and waypoint information in a heading-up mode. Distance scaling was deactivated to allow for identical presentation to all participants.

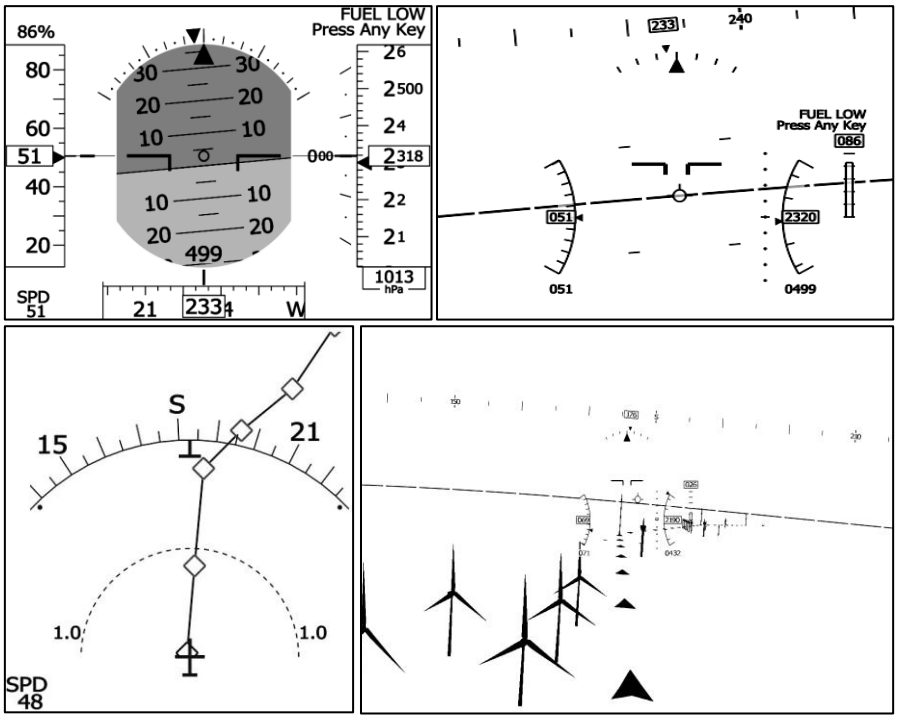


Figure 2. Head-down displays (left) and helmet-mounted display (right). Primary flight displays (top) and navigation displays (bottom).

Scenario design

Scenarios were started from a freeze position in the air. Each scenario consisted of an en route and an approach/landing phase and took approximately 10-12 minutes to complete. Two visibility conditions occurred and visibility always changed after half of the en route phase. Poor visibility provided a visual range of approximately 800 m, and average visibility provided a range of 1200 m.

Expected events

During the en route segment pilots were instructed to perform two different tasks, a monitoring task on the display, and a search and identification task in the outside scene. Within the monitoring task they were briefed to monitor for heading, speed and altitude changes and adjust the parameters timely and accordingly. Twelve

changes occurred in each scenario. In the far domain task they were instructed to search for small fuel trucks positioned in the terrain. Trucks differed in their colour: green trucks were labelled as friend targets, red as foe targets, and grey as neutral targets. Participants had to detect the trucks, determine the correct colour and report it by pushing a corresponding button on the centre stick. The particulars of the main tasks are presented in detail in Knabl et al. (2014).

Unexpected events

Additionally two unexpected events occurred, one on the display and one in the outside scene. Pilots were not briefed about the appearance of these events. Both appeared twice, once with the HMD, and once with the baseline condition. Thus each event was only once truly unexpected. For half of the participants the events first occurred with the HMD and afterwards with the baseline condition, for the other half it was the opposite. Furthermore, for one half they appeared in poor, for the other in average visibility.

The display event consisted of a warning stating “fuel low press any key”. The warning appeared above the altitude tape, blinked with 2 Hz for four seconds and then remained steady until any button was pushed (figure 2, top). The far domain event consisted of an intruder helicopter hovering in the flight route (figure 3). Pilots were required to detect the aircraft and perform an adequate collision avoidance manoeuvre. The helicopter was only visible in the outside scene since no traffic was presented on any display.

Figure 3. Far domain event: intruder helicopter hovering in flight route

Data analysis

Data were analysed with SPSS Statistics 20. An alpha level of .05 was adopted for significance. Statistical analysis was conducted using repeated measures analysis of variance (ANOVAs) and t-tests. Data are further presented as mean (M) and standard deviation (SD).

Results

Helicopter event detection

Firstly, the frequency of lateral and vertical manoeuvres was determined to assess the overall quality of collision avoidance. Lateral manoeuvres, especially right turns, were regarded as most adequate. Based on pilot comments, a right turn would be the

typically used manoeuvre (in countries with right-hand traffic), although it is not specifically stated as a rule. Therefore left turns were considered as adequate as well, given that the pilot recognised the helicopter being in a hover and not in a forward movement. Vertical manoeuvres however were regarded as less appropriate, since the helicopter would either receive or cause turbulence due to the rotor downwash. Descriptive results indicated that the helicopter was predominantly avoided by a lateral manoeuvre and right turns were also most frequently selected. A vertical manoeuvre was selected only twice with the PFD condition, however six times with the HMD. In one PFD scenario the pilot was so far off-track that no reaction was required. Finally, one pilot in the HMD condition did not react at all and commented that he would have probably collided with the helicopter, if he had not already been at a too high altitude. Apart from that near-miss, no collision occurred.

Table 2. Descriptive results of collision avoidance manoeuvre type

Collision avoidance manoeuvre type	frequency		minimal distance achieved (m)	
	HMD	PFD	HMD	PFD
Right turn (behind helicopter)	6	9	33.5	24.6
Left turn (in front of helicopter)	5	6	32.9	25.7
Descent (below helicopter)	2	2	13.3	13.4
Climb (above helicopter)	4	-	16.3	-
No reaction (off track)	-	1	-	-
No reaction (not detected)	1	-	-	-

Secondly, based on the visual inspection of the manoeuvre and analysis of the control inputs, the start of the avoidance manoeuvre was determined and the distance to the helicopter was calculated. However, it has to be noted that the exact starting point was not always apparent. Therefore statistical analysis was carried out only for 12 pilots. Due to the rather small sample size visibility and order of appearance were not accounted for. A one-way repeated measures ANOVA revealed no significant main effect of display condition, $F(1, 11) = 1.41$, $p = .260$, $\eta^2_p = .11$. Thus the distance to the helicopter at start of avoidance did not statistically differ between the HMD (M = 366.5, SA = 144.7) and the baseline (M = 444.3, SA = 180.8).

Warning detection

Reaction time from warning appearance to response was calculated as a function of display type (HMD/baseline), visibility condition (poor/average) and order of appearance (HMD or baseline first). A three-way split-plot ANOVA was calculated with the between subject factors visibility and order, and the within subject factor display. No significant main effects were obtained. However results revealed a significant interaction of display x order, $F(1, 12) = 7.0$, $p = .021$, $\eta^2p = .369$. Post-hoc t-tests for independent samples were calculated and revealed a significant order effect only for the HMD, $t(15) = 2.6$, $p = .020$, but not for the baseline, $t(15) = -1.2$,

p = .248. Hence, when the warning appeared on the PFD, reaction time did not differ as a function of order, thus whether it occurred for the first (M = 3.2, SA = 1.0) or for the second time (M = 2.7, SA = 0.9). However, it was found that when the warning was presented on the HMD, pilots responded significantly later if it was truly unexpected (first: M = 4.5, SA = 1.6; second: M = 2.6, SA = 0.9). Moreover, the second order interaction (display x visibility x time) was also found to be significant, $F(1, 12) = 11.5$, $p = .005$, $\eta^2p = .489$. As illustrated in figure 4, the finding strongly indicated that the HMD reaction cost to the truly unexpected warning was only apparent in poor (M = 5.4, SD = 0.8) but not in average visibility (M = 3.0, SD = 1.3).

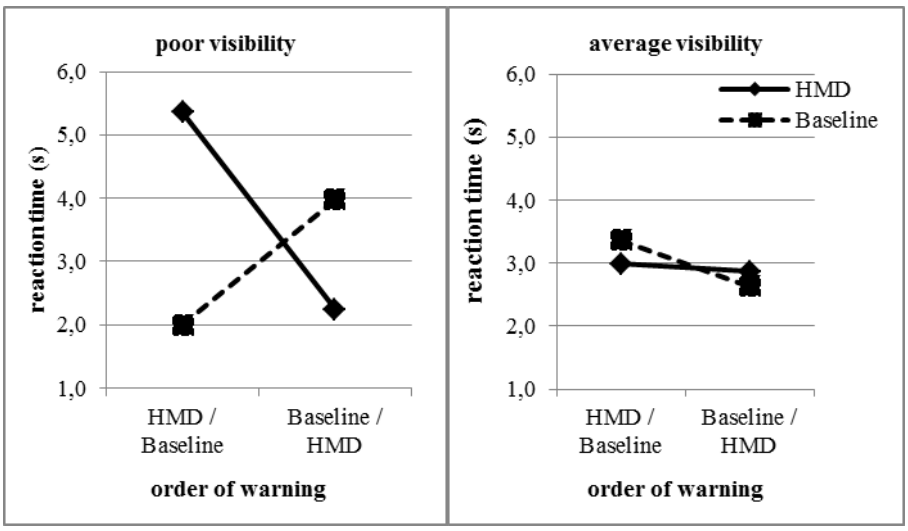


Figure 4. Reaction time (s) to warning as a function of display type, visibility and order

Discussion

With regard to the fuel low warning on the display, results revealed a longer reaction time with the HMD if the warning is truly unexpected. Interestingly the cost is only apparent in poor visibility, but not in average. It is assumed that in poor visibility pilots directed more attention to the far domain in order to avoid obstacles and search for targets, therefore the warning was responded to later. In contrast, the search task was less time-critical during average visibility conditions, enabling a more successful division of attention between the two domains and therefore a faster detection of the warning. However, it has to be assumed that attention was also more focused on the outside scene during the poor visibility PFD condition, although the detection cost is not apparent here. One possible reason for this is that the warning on the PFD - while focusing outwards - was presented in the peripheral visual field, which contains a large number of rods and is associated with a higher sensitivity compared to the fovea (Goldstein, 2013). Moreover the detection drawback was only obtained for the truly unexpected warning and disappeared with the second presentation. Thus saliency alone did not favour rapid detection, but expectancy did;

a finding which is in accordance with the SEEV model. Selective attention could therefore be directed quickly to the warning on the HMD, when it was presented for the second time. Further investigations should also raise the question whether expectancy is dependent on the location of the warning. Hence, is reaction time only reduced when the second warning appears at exactly the same position, or are test subjects also more susceptible to unexpected display events presented at a different display location?

With regard to the helicopter event, the descriptive results indicate that predominantly the most appropriate, lateral manoeuvre was chosen. However, vertical manoeuvres were selected more frequently with the HMD (six times) than with the baseline (twice). It has to be noted that no statistical analysis of the frequencies was performed due to the very small group size. No significant differences were obtained for the distance-based evaluation, indicating that pilots did not start their avoidance manoeuvre later with the HMD. Nevertheless, the results in general indicate a very slight but consistent tendency towards an HMD drawback that is supported by the following considerations. First, the higher frequency of vertical manoeuvres, second, the indication that descriptively pilots started the avoidance manoeuvres later. Third, at least one pilot specifically commented on indeed detecting the helicopter, however not having had enough time and cognitive resources left to consider a proper avoidance plan, which again might somewhat explain the higher frequency of vertical manoeuvres. Finally, the fact that one pilot did not detect it at all is consistent with findings from head-up display literature and is attributed to both the effect of clutter and attention fixation.

To sum up, the present paper provides evidence that, under certain conditions, HMD pose a risk of inducing event detection costs and that these hold true for both, events on the display and in the far domain. The findings are therefore consistent with previous results obtained from head-up display experiments. To mitigate these costs, technology-based solutions as well as human-centred solutions should be accounted for. With regard to technology, the implementation of enhanced vision based on real-time sensor data is a key factor. Highlighting or cueing objects such as traffic or obstacles on the HMD provide the possibility to specifically direct attention to these hazards. In addition, it is of interest whether detection performance can be improved by proper training, which would address the vulnerability to inattentional blindness and attentional capture and would create awareness of the susceptibility to these effects. Finally, for dual pilot operations, research should focus on crew procedures, task sharing and management as well as team situation awareness as well.

Acknowledgements

This work was partly sponsored by the German Department of Defense (DOD) within different projects. The authors especially would like to thank all participants for their support and their valuable contributions.

References

Eriksen, C.W., & Eriksen, B.A. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics, 16*, 143-149.

Fadden, S., Ververs, P.M., & Wickens, C.D. (1998). Costs and benefits of head-up display use: A meta-analytic approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 42*, 16-20.

Goldstein, E. B. (2013). Sensation and Perception, Cengage Learning.

Knabl, P., Schmerwitz, S., Doehler, H.-U., Peinecke, N., & Vollrath, M. (2014). *Attentional issues with helmet-mounted displays in poor visibility helicopter flight*. Paper presented at the 31st EAAP Conference, Valetta, Malta.

Lorenz, B., Többen, H., & Schmerwitz, S. (2005). Human performance evaluation of a pathway HMD. *Proceedings SPIE 5802, Enhanced and Synthetic Vision*, 166-176.

Mack, A. (2003). Inattentional Blindness: Looking Without Seeing. *Current Directions in Psychological Science, 12*, 180-184.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. Cambridge, MA, US: The MIT Press.

Wickens, C.D., & Alexander, A.L. (2009). Attentional Tunneling and Task Management in Synthetic Vision Displays. *The International Journal of Aviation Psychology, 19*, 182-199.

Wickens, C.D., Helleberg, J., Goh, J., Xu, X., & Horrey, W.J. (2001). *Pilot task management: Testing an attentional expected value model of visual scanning (ARL-01--14/NASA-01--7).* Savoy, IL*:* University of Illinois, Aviation Research Lab.

Wickens, C.D., & Long, J. (1994). Conformal Symbology, Attention Shifts, and the Head-Up Display. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 38*, 6-10.

Wickens, C.D., & Long, J. (1995). Object versus space-based models of visual attention: Implications for the design of head-up displays. *Journal of Experimental Psychology: Applied, 1*, 179-193.

Wickens, C.D., & McCarley, J.S. (2007). *Applied Attention Theory*. Boca Raton, FL: CRC Press.

A novel Human Machine Interaction (HMI) design/ evaluation approach supporting the advancement of improved automation concepts to enhance flight safety

Joan Cahill & Tiziana C. Callari
Centre for Innovative Human Systems (CIHS),
School of Psychology, Trinity College Dublin,
Ireland

Abstract

This paper presents a novel Human Machine Interaction (HMI) design/evaluation methodology, supporting the specification and evaluation of a new adaptive automation concept, both from a functional and an operational/safety perspective. This methodology has been advanced as part of the work requirements for the Applying Pilot Models for Safety Aircraft (A-PiMod) project, funded by the European Commission. Critically, this methodology integrates/combines formal HMI design/evaluation approaches (i.e. user interviews and simulator evaluation) with an integrated stakeholder approach to evaluation. The objective of this paper is to highlight (1) what is new in this overall approach (i.e. integration of formal HMI approaches such as simulator evaluation with stakeholder evaluation approaches, decomposition of project goals to project objectives, evaluation objectives and key performance indicators); (2) what is new in the specific stakeholder approach to evaluation (i.e. the set-up of a Community of Practice involving both internal and external stakeholders, and the integration of this methodology with wider HMI evaluation activities); and (3), what the methodology delivers in terms of ensuring improved levels of safety and reliability for the aviation sector. The evaluation of this methodology will be based on an analysis of project outcomes to date.

Introduction

The air accident and flight safety literature reports on the many still-open human factors issues concerning automation design. For example: Flight Air France 447 (2009), Flight Spanair 5022 (2008), Flight Helios Airways HCY 522 (2005), Flight China Airlines 140 (1994), and Flight Air Inter 148 (1992).

Several human factors problems have been documented in relation to automation design. This includes: automation surprises (i.e. the crew does not understand what automation is [or is not] doing), workload concerns (i.e. whether or not automation actually increases workload in certain situations, given that the crew have to track the status/actions of automation, and/or lack of workload support in high workload situations), and issues pertaining to over-reliance on automation (i.e. potential that over reliance on automation might have a negative impact on pilot flight

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

management skills/competencies, overall impact on expertise etc.). In addition, certain aspects of automation design require more detailed consideration. Currently, the dynamic task allocation between the crew and automation is based on an assessment of the aircraft state (i.e. aircraft systems only). Indeed, automation is not really aware of the crew and at times, it acts forcefully. In this sense, new automation concepts must address the issues of teamwork (i.e. how to support co-operation/teamwork, what aspects of crew state to consider and how to distribute workload/tasks between the crew and automation). Also, some key questions concerning automation and the role of the pilot have not been fully addressed (i.e. level of authority in relation to key flight management tasks and vetoing automation). These questions can also be posited from an automation perspective (i.e. can or should automation veto the pilot's decision?). In this regard, existing automation systems have built in 'protections' to ensure that the aircraft remains in a safe state. This mainly concerns abnormal 'safety critical' situations. Critically, the design of an improved automation system should support pilot task performance – and specifically, address the issues as outlined above.

This paper reports on a new Human Machine Interaction (HMI) design/evaluation methodology supporting the specification and evaluation of a new adaptive automation concept, both from a functional and an operational/safety perspective. This methodology has been advanced as part of the work requirements for the Applying Pilot Models for Safety Aircraft (A-PiMod) project, funded by the European Commission. First, a short introduction to the A-PiMod project and the Researcher's role in the project (i.e. Human Factors research team from Trinity College Dublin) is provided. Existing HMI design/evaluation methods are then reviewed. Following this, the proposed novel HMI methodology is presented. An overview of the specific validation activities designed and implemented to date is then reported. Following this, the main outcomes and project achievements are reviewed. The benefits and application of this approach is then discussed. Finally, some conclusions are drawn.

The Applying Pilot Models for Safety Aircraft (A-PiMod) project

The A-PiMod project aims to address certain still-open automation problems, as outlined above. The high level goal of the project is to improve flight safety in a time of increasing levels of performance, automation and information provision to the flight deck. Specifically, the objective of the A-PiMod Project is to design a new adaptive automation concept based on a hybrid of three elements – (1) Multi-Modal Pilot Interaction, (2) Operator Modeling, and (3) Real-Time Risk Assessment. Three impact statements have been defined to assess the expected project outcomes: (1) to reduce accident rate by 80%; (2) to achieve a substantial improvement in the elimination of and recovery from human error; (3) to mitigate the consequences of survivable accidents.

The high level objective of our research in this project is to validate the A-PiMod concepts and technologies from a (1) functional and, (2) operational/safety perspective. This spans requirements specification/validation, prototype design and evaluation, and the final evaluation of safety/operational impact. To do so, a novel

methodology has been proposed to support the specification and evaluation of the new adaptive automation concept. This is discussed in a later section.

Overview of existing HMI approaches to evaluation

The HMI literature defines a range of formal and informal methods for the design of human friendly technology adopting a 'User-Centered Design' methodology (Cooper, 2007; Preece, Rogers, & Sharp, 2007; Constantine & Lockwood, 1999; Hackos & Reddish, 1998). The specific approaches adopted reflect underlying theoretical assumptions about design practice. In particular, they represent diverse views concerning the role of end users, the specific process for envisioning new technology requirements, and the relationships between design and evaluation.

Formal HMI Design/Evaluation Methods

Typically, formal HMI methods start with analysing the existing task (Preece et al., 2007). To this end, a task analysis is first undertaken, involving the participation of end users. Structured or semi-structured interviews are used to understand and evaluate current work practices and supporting technology requirements (Hackos & Redish, 1998). Several analysis steps are then undertaken without the participation of end users. Analysis outputs include lists of end users, user and task matrices and task workflow diagrams. This is followed by different design activities such as storyboarding and prototyping. Once the prototype is developed, users are involved in different evaluation activities. In this way, design and evaluation are conceived as separate steps.

Informal HMI Design/Evaluation Methods

Formal HMI methods have been the subject of much debate in the HCI literature. Specific challenges have come from the fields of Ethnography and Participatory Design. Ethnographers argue that classical HCI methods do not take work practice seriously; failing to address the social aspects of work (Hutchins 1995; Vicente 1999). Participatory design theorists have questioned the separation between design and evaluation in formal methods (Bødker & Buur, 2002). Specifically, they have challenged the instructiveness of traditional user and task analysis outputs for design guidance. Central to Participatory Design theory is the idea that Usability Engineers design 'with' end users, as opposed to 'for' them. Accordingly, users are active participants in the design process (Bannon & Bødker, 1991, Bødker & Grønbæk, 1996). Several techniques are outlined in the literature. This includes concept generation, envisionment exercises, scenario role playing, story collecting and storytelling (through text, photography and drama), and the co-creation and evaluation of prototypes.

Operational Validation/Evaluation approaches

Arguably, existing HMI design/evaluation methods fail to address the broader operational issues underpinning the envisionment and specification of new technologies. Operational assessment involves more than the assessment of operator performance (i.e. in relation to task workflows, workload and situation awareness),

and the allied performance of the proposed system (i.e. usability of the proposed system/user interface). Crucially, wider 'operational' issues must be considered. This includes the fit between the technologies and the proposed operational scenarios, the specification of operational requirements (at a process as well as a task level), the assessment of operational benefits, the design of future operational processes/procedures, the specification of teamwork/co-ordination and information sharing requirements across relevant system actors, and the identification of potential implementation barriers.

Stakeholder approaches to evaluation

The involvement of stakeholders as part of programme/project evaluation has received increasing attention over the past three decades (Rodriguez-Campos, 2011). Overall these approaches follow from the idea that collaboration must tackle issues that matter and have impact/benefits for the stakeholder's organization/domain of activity. Further, such collaboration requires a high level of interpersonal and organizational trust. Central to this, is the establishment of communication and discussion methods/sessions. The use of knowledge generation and tacit knowledge elicitation methods are favoured in these approaches. These methods promote ways to transfer users' tacit knowledge as a source of sustainable competitive advantage. Stakeholder evaluation approaches do not necessarily involve technology design/evaluation. For example, such approaches have been applied to the evaluation of processes, the delivery of services, events, architecture, the layout of cities and relevant social spaces (i.e. parks/playgrounds), and so forth.

The novel HMI Design/Evaluation Methodology adopted in A-PiMod

Introduction to Research

The validation activities will address the following key issues pertaining to automation design:

- The design of the cockpit as a co-operative system (i.e. Pilot/automation co-ordination/teamwork, distribution of task activity between the crew and automation);
- Pilot comprehension of automation (i.e. status of automation, who is responsible for what task and what are they doing) and the avoidance of automation surprises
- How automation might be designed to enable workload management and reduce crew stress in high workload and potentially safety critical situations;
- How the A-PiMod concept enables/supports crew briefing/planning, situation assessment, information management and decision making (linking to Crew Resource Management concepts);
- How the A-PiMod concept enables/supports error identification and recovery.

Overall, the evaluation approach involves two strands of activity – (1) research with the A-PiMod Community of Practice, and (2) formal simulator evaluation. Collectively, this research can be characterized in relation to two key features - (1) early design/evaluation and (2) iterative design/evaluation.

Validation activities in A-PiMod are designed to be both early and iterative. Validation occurs after the initial specification of requirements elicitation and review (milestone 1), and then at two key milestones in project (milestone 2 and milestone 3). The first round of simulator evaluations (i.e. validation cycle 1/milestone 2) are designed to be explorative (i.e. using low fidelity prototypes), while the second round (i.e. validation cycle 2/milestone 3) will involve a full scenario run (i.e. using high fidelity prototypes). Also, there is on-going validation with internal and external stakeholders. Further, there will be a final evaluation of the overall system in relation to the overall safety/operational impact (i.e. milestone 4). For a graphical illustration of this, please see Figure 1 below.

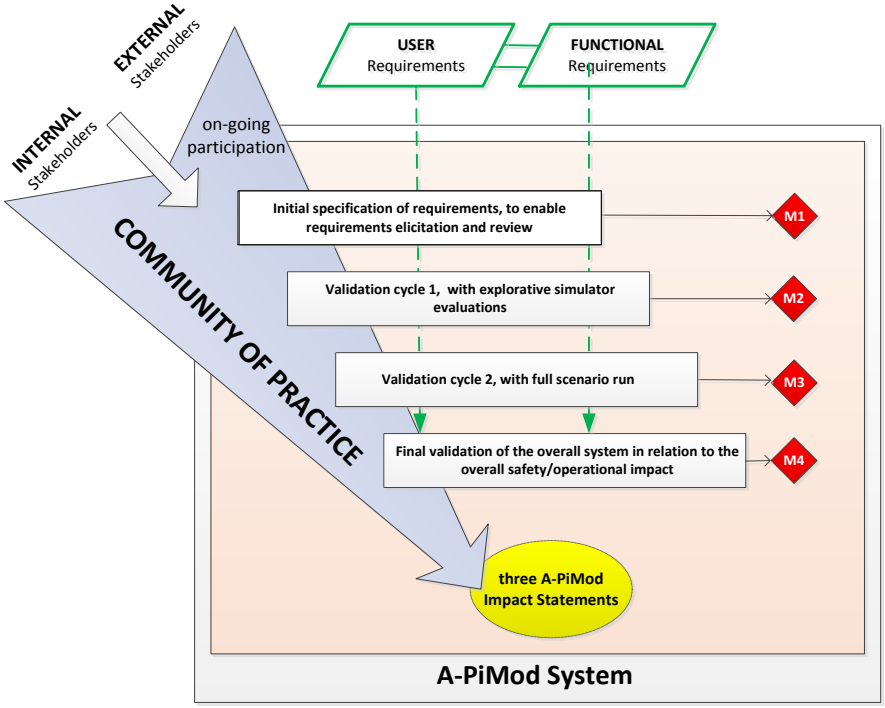


Figure 1. Validation timeline/activities in A-PiMod.

The methodology proposed in A-PiMod integrates/combines formal and informal HMI design/evaluation approaches, along with an integrated stakeholder approach to evaluation. Overall this is an iterative process and links to the documentation of functional/technical requirements and associated prototyping activities.

The following sections will outline what is new and/or innovative in the proposed methodology, in relation to the following perspectives:

(1) What is new in the overall validation approach;
(2) What is new in the specific stakeholder approach to evaluation;
(3) What the methodology delivers in terms of ensuring improved levels of safety and reliability for the aviation sector.

New: The overall validation approach

The validation process in A-PiMod will support the assessment of how far the proposed technologies achieve the project goals and outcomes. It is underpinned by (1) User-Centred Design concepts and methods, and (2) the principle that safety is and operational concept. In determining the project evaluation objectives and questions, a hierarchical decomposition has been followed to ensure that validation activities are focussed on project outcomes and provides traceability. This process starts from the A-PiMod project goals/objectives, through to evaluation objectives (high level and detailed), evaluation questions (high level and detailed, and key performance indicators (KPI).

The proposed methodology integrates formal HMI approaches (such as simulator evaluation), with informal/participatory HMI methods (such as collaborative prototyping), along with tacit knowledge elicitation methods (such as semi-structured interviews following specific techniques – i.e. the Critical Incident Technique (Butterfield, Borgen, Amundson, & Maglio, 2005; Flanagan, 1954) and the Instructions to the Double technique (Oddone & Re, 1994; Oddone, Re, & Briante, 2008; Re & Oddone, 1991)).

New: The specific stakeholder approach to evaluation

Validation activities in A-PiMod have involved the application of a participatory/stakeholder approach to evaluation. The stakeholders involved in A-PiMod are referred to as the A-PiMod Community of Practice. Critically, these activities have developed a working collaboration with experts, which includes both 'primary users' (i.e. *internal stakeholders* representative of each project partner) and 'all legitimate groups' (i.e. *external stakeholders* representative of the aviation-related industry and Flight operational system). Both sets of stakeholders are involved in the specification and evaluation of the emerging adaptive automation concepts and technologies. This spans several activities pertaining to the specification and evaluation of user/technical requirements and user interface design prototypes. Internal stakeholders provide input based on their own domain knowledge. Further, they contribute in relation to assessing what is technically feasible and possible from a project perspective. On the other hand, external stakeholders provide feedback from direct experience and practice, to ensure that the emerging solution addresses real operational and safety requirements. Both internal and external stakeholders are conceived as active collaborators and contribute/engage in validation exercises on an on-going basis.

In the validation activities with the A-PiMod Community of Practice TCD's role goes beyond that of a neutral facilitator. TCD's role is to actively promote an interactive learning environment, where the stakeholders share their expertise and learn from the group collaboration. Indeed, TCD also act as a 'key-broker role' between the members of the Community of Practice to support (1) the review and specification of user requirements for the future system, (2) the production of relevant user interface design concepts/prototypes, and (3) the evaluation of prototypes.

New: What the methodology delivers in terms of ensuring improved levels of safety and reliability for the aviation sector.

A safety case has been advanced to support the specification of requirements and the assessment of safety/operational impact. The safety case comprises two parts – (1) the theoretical framework for the safety case and (2) the specific safety argument.

The safety framework provides a principled basis for conceptualizing/demonstrating how the A-PiMod adaptive automation concept and associated technologies will yield specific operational and safety benefits. This links to the demonstration of project impact, as discussed earlier. The framework is reported as a progression of ideas which form several phases. This includes: (1) background concepts which underpin the safety framework, (2) the starting point for conceptualizing the safety case, (3) the A-PiMod concept, and (4) the benefits of the A-PiMod adaptive automation concept and associated technologies from an operational and safety perspective. Each phase is associated with key points. The overall framework is depicted in Figure 2.

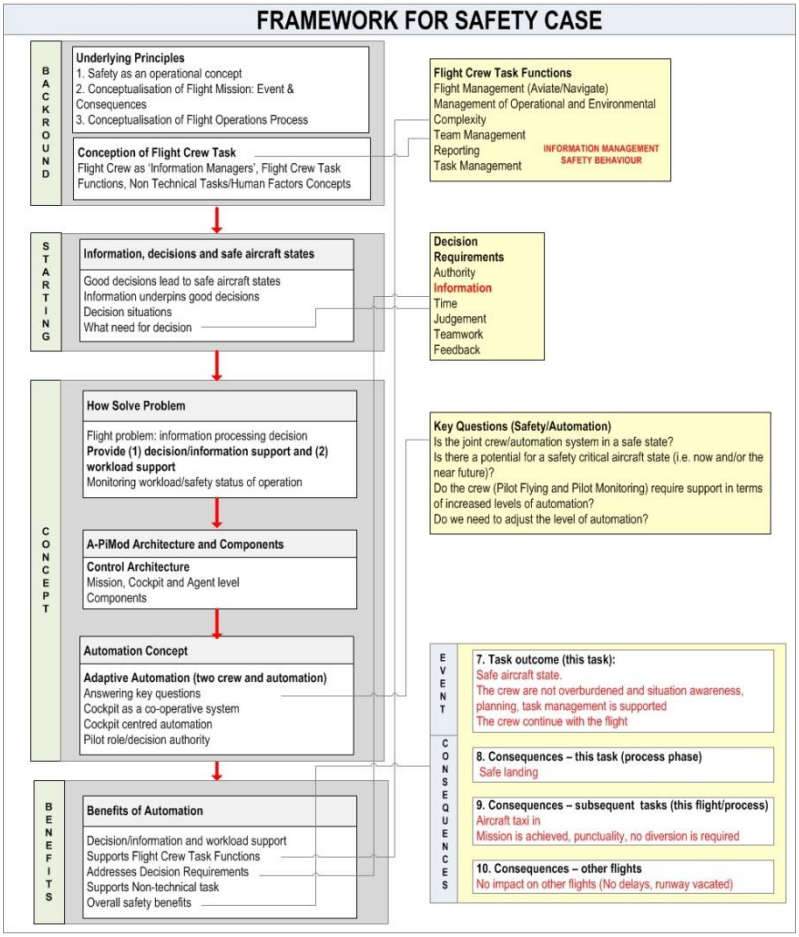


Figure 2. Safety Framework.

The safety argument articulates how specific operational/safety goals are achieved at the level of the A-PiMod technology (i.e. proposed architecture and technical components). Overall, the argument structure follows the theoretical approach and specific automation concept, as outlined in the safety framework. Specifically, the safety case/argument refers to specific steps in an overall use scenario – i.e. what technology does at different points in the scenario. See Figure 3 below.

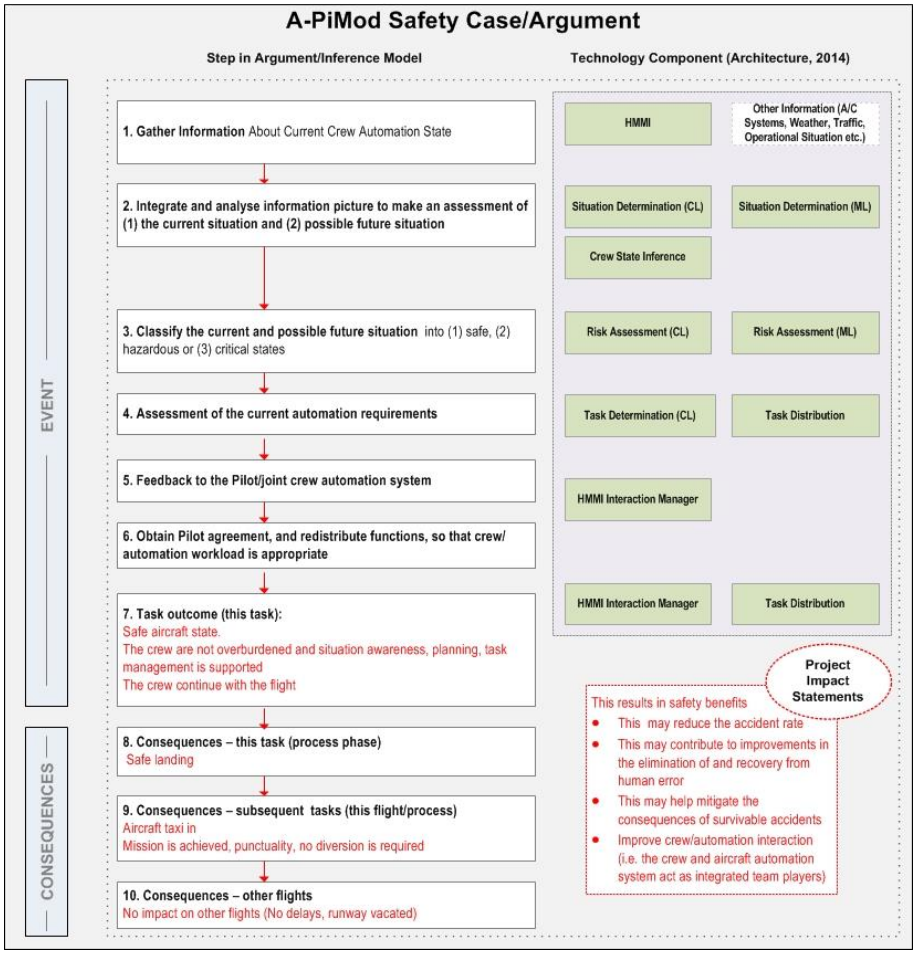


Figure 3. Safety argument.

What has been achieved so far in terms of validation approach

Research Undertaken

The project started in September 2013. Overall, this research has involved two strands of activity – namely, (1) ongoing validation research with the Community of

Practice, and (2) the preparation of Validation Cycle 1 (i.e. comprising simulator evaluation, a parallel desktop evaluation and training evaluation).

In relation to (1), the on-going research activities with the A-PiMod Community of Practice, eight validation exercises sessions involving both internal and external stakeholders have been implemented. Overall, the objective of these sessions was to (1) define and integrate the A-PiMod adaptive automation concept, and (2) to harmonise the allied user/functional requirements. Both remote (by means of the WebEx functionality) and face-to-face workshops and/or interviews were undertaken. Prior to the validation exercise workshops, members of the Community of Practice were asked to complete tasks as defined by TCD. This served to facilitate the learning environment and promote the sharing of ideas and discussion in the specific workshops and/or interview sessions. Following every validation exercise workshop, TCD reported the minutes of the workshop and the consensus obtained on the topic. Further, TCD designed session specific templates to highlight the main results and integration of the Community of Practice members' feedback.

In relation to (2), the first formal validation of the A-PiMod concept will take place in November 2014. The first Validation Cycle aims to evaluate and further specify the A-PiMod (1) adaptive automation concept, (2) the Multi-Modal Interaction concept and, (3) the training concept. In relation to the A-PiMod (1) adaptive automation concept, and (2) the Multi-Modal Interaction concept, this will involve an explorative user test with Pilots (i.e. four sets of crew), using a simulator. In addition, there will be some parallel evaluations (i.e. outside the simulator) with the same panel of Pilots (i.e. participatory review/design of concepts, semi-structured interviews to evaluate the concepts and so forth). In relation to (3), this will involve a parallel evaluation of the training concept, using semi-structured interviews.

The validation activities have produced a huge amount of qualitative data. Data recording and analysis has been undertaken with the assistance of a Computer-Assisted Qualitative Data Analysis Software (CAQDAS) tool - NVivo (© QSR International, V.8) (Bazeley, 2007). The use of difference sources of evidence during the data collection (i.e. interviews, observations, collaborative prototyping, etc.) allows for the assessment of convergence in relation to data evidence (data triangulation). This contributes to research validity. Further, the use of a concept-driven coding frame (based on the architecture and technology that A-PiMod intends to demonstrate) has supported the ongoing data analysis.

Emerging A-PiMod Adaptive Automation Concept

This research (i.e. use of innovate HCI design/evaluation methodologies) has resulted in the specification of (1) a new adaptive automation concept/approach and (2) the associated new technology concepts and requirements.

The problem of flying the aircraft is conceptualised as an 'information processing decision'. This can be achieved in different ways (i.e. two/one person cockpit with different levels of automation, ground co-Pilot and/or ground support, or UAV/drone). In A-PiMod, these decisions will be undertaken by a two person crew with the support of automation. This is referred to as a 'co-operative system'. The

underlying idea is that we can continuously monitor the operational situation and the allied crew/automation state, to determine the best distribution of task activity between the crew and automation. The basic philosophy is - if there is an increase in workload, certain functions can be shifted to automation, to reduce the burden on the flight crew. Automation is also used to support information management and decision making tasks.

Critically, the A-PiMod system allows us to answer the following questions:

- Is the joint crew/automation system in a safe state (i.e. level of workload, situation awareness)?
- Is there a potential for a safety critical aircraft state (i.e. now and/or the near future)?
- Do we need to adjust the level of automation?

The crew obtain constant feedback via a new cockpit user interface as to status of (1) the operational situation, and (2) the joint crew automation system. From an operational/safety perspective this enables crew/automation teamwork, crew workload management, and error identification and recovery. All of the above ensures that the aircraft remains in a safe state. This in turn has consequences in relation to the overall safety of the flight, and the achievement of process/operational goals.

Discussion

The integration of formal and informal HMI methods, along with a stakeholder approach to evaluation has proved effective in relation to the specification of the A-PiMod concept. As outlined above, this has resulted in the preliminary advancement of an innovative approach to automation, which addresses known problems.

Several points in relation to the stakeholder approach to evaluation should be noted. First, the implementation of Community of Practice research is not straightforward. This requires the advancement of a ‘working relationship’ with community members (i.e. trust and teamwork), the set-up and acceptance of communication/information sharing practices and the establishment of a decision making process. All of this takes time. Further, the adoption of a participatory approach can make decision making slow. However, on the positive side, this in turn fosters collaboration and good co-ordination across project members.

In this regard, the TCD role has changed over the course of these validation activities. Initially our role was one of a ‘facilitator’ and/or coordinator. We sought to capture requirements and to advocate on behalf of the end user. Over time, we have become more and more engaged in the current implementation of project activities (i.e. in eliciting Human Factors requirements, suggesting user requirements, designing user interface prototypes and so forth). In doing this, TCD has adopted a ‘brokerage role’ between internal/external stakeholders. This is underpinned by quality communication and the establishment of good working relationships between TCD and internal/external stakeholders (i.e. trust and teamwork).

The creation of an inclusive learning environment where members of the A-PiMod Community of Practice share ideas necessitates an appropriate setting (and potentially technology). In A-PiMod this has been mostly remotely telephone/web mediated (i.e. with WebEx), although some person to person interviews have been undertaken. Overall, person-to-person interaction has proved the most fruitful. As a result, the planning of the next validation exercises will consider more opportunities to meet in person. In time, technology may 'catch up', to provide a more natural/user-friendly environment for knowledge sharing.

Lastly, the importance of involving external stakeholders (i.e. pilots) cannot be understated. This involvement has been critical to the collection of user requirements and the emerging definition of the A-PiMod concept.

Conclusions

Safety is an operational concept and must be addressed at all levels: the air traffic management (ATM) system; the design of airline safety management system (SMS) processes and technologies; flight crew task activities and in particular, flight crew safety behaviour, and the design of cockpit systems/tools (including automation).

Overall, the evaluation/validation approach adopted has facilitated the preliminary specification and evaluation of a new adaptive automation concept. Specifically, the integration of a range of formal and informal HMI methods has proved effective in terms of enabling both operational and safety validation. The participation of stakeholders in the Community of Practice provides a strong link to the real world – in relation to (1) understanding automation issues, and (2) the capacity of technology to address these issues. Critically, the emerging adaptive automation concept is predicated on feedback in relation to flight crew experience with automation (and associated problems).

It is anticipated that these initial concepts will pave the way for an improved approach to automation. Preliminary evaluation feedback indicates that the concepts/technologies show promise in relation to solving pilot problems relating to teamwork (i.e. pilot/automation co-ordination) and workload management.

Acknowledgements

The research leading to these results/preliminary outcomes has received funding from the European Commission's Seventh Framework Programme (FP7/2007-2013) under grant agreement N. 605141 - Applying Pilot Models for Safety Aircraft (A-PiMod) Project.

Thanks to the project coordinator – Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) and to the other project partners – OFFIS (OFF), Honeywell (HON), National Aerospace Laboratory of the Netherlands (NLR), Symbio (SYM), Kite Solutions (KITE), Brno University of Technology (BUT). Also, thanks to the members of the A-PiMod Community of Practice, in particular the external stakeholders (i.e. pilots).

References

Bannon, L., & Bødker, S. (1991). Beyond the Interface: Encountering Artefacts in Use. In J.M. Carroll (Eds.), *Designing Interaction: Psychology at the Human-Computer Interface* (pp. 227-253). New York: Cambridge University Press.

Bazeley, P. (2007). *Qualitative Data Analysis with NVivo*. London: SAGE Publications.

Bødker, S., & Burr, J. (2002). The Design Collaboratorium. A Place for Usability Design. *ACM Transactions on Computer Human Interaction, 9*(2), 152-169.

Bødker, S., & Grønbæk, K. (1996). Users and Designers in Mutual Activity. An analysis of cooperative activities in systems design. In Y. Engeström and D. Middleton (Eds.), *Cognition and Communication at Work* (pp. 130-158). Cambridge: Cambridge University Press.

Butterfield, L.D., Borgen, W.A., Amundson, N.E., & Maglio, A.S.T. (2005). Fifty years of the critical incident technique: 1954-2003 and beyond. *Qualitative Research, 5*(4), 475-497.

Constantine, L.L., & Lockwood, L.A.D. (1999). *Software for Use: A Practical Guide to the Models and Methods of Usage-centered Design*. MA: Addison-Wesley.

Cooper, A. (2007). *The inmates are running the asylum* (7th ed.). Indianapolis: SAMS Publishing.

Flanagan, J. C. (1954). The Critical Incident Technique. *Psychological Bulletin, 51*, 327-358.

Flight Air Inter 148 (1992, January 20). *Final Report on the accident on 20th January 1992 involving an Airbus A320 registration F-GGED operated by Air Inter Airlines in Vosges Mountains (near Mont Sainte-Odile).* Retrieved from Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (BEA). http://www.bea.aero/docspa/1992/f-ed920120/htm/f-ed920120.html

Flight China Airlines 140 (1994, April 26). *Final Report on the accident on 26th April 1994 involving an Airbus Industrie A300B4-662R registration B1816 operated by China Airlines at Nagoya Airport (Published July 19, 1996).* Retrieved from Aircraft Accident Investigation Commission. http://www.skybrary.aero/bookshelf/books/808.pdf

Flight Helios Airways HCY522 (2005, August 14). *Final Report on the accident on 14th August 2005 involving a Boeing 737-31S registration 5B-DBY operated by Helios Airways at Grammatiko, Hellas (Published November 2006).* Retrieved from Air Accident Investigation & Aviation Safety Board (AAIASB). http://www.moi.gov.cy/moi/pio/pio.nsf/All/F15FBD7320037284C2257204002B6243/$file/FINAL%20REPORT%205B-DBY.pdf

Flight Spainair 5022 (2008, August 20). *Final Report on the accident on 20th August 2008 involving a McDonnell Douglas DC-9-82 (MD-82) registration EC-HFP operated by Spainair at Madrid-Barajas Airport (Published October 8, 2008).* Retrieved from Comisión Investigatión de Accidentes e Incidentes de Aviación Civil (CIAIAC). http://www.fomento.es/NR/rdonlyres/EC47A855-B098-409E-B4C8-9A6DD0D0969F/107087/2008_032_A_ENG.pdf

Flight AF 447 (2009, June 1). *Final Report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro (Published July 2012)*. Retrieved from Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile (BEA). http://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf

Hackos, J.A., & Redish, J.C. (1998). *User and Task Analysis for Interface Design*. New York: Wiley Computer Publishing, John Wiley & Sons Inc.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Oddone, I., & Re, A. (1994). Come recuperare l'expertise professionale. In P. Tranchina (Ed.), *Portolano di Psicologia* (pp. 277-280). Pistoia: Centro di Documentazione Pistoia.

Oddone, I., Re, A., & Briante, G. (2008). *Esperienza operaia, coscienza di classe e psicologia del lavoro*. Torino: OTTO Editore.

Preece, J., Rogers, Y., & Sharp, H. (2007). *Interaction Design: Beyond Human-Computer Interaction* (2nd ed.). West Sussex, UK: John Wiley & Sons Ltd.

Re, A., & Oddone, I. (1991). Competenza professionale ristretta e competenza professionale allargata. In G. P. Quaglino (Ed.), *Soggetti lavoro professioni* (pp. 260-274). Torino: Bollati Boringhieri.

Rodriguez-Campos, L. (2011). Stakeholder Involvement in Evaluation: Three Decades of the American Journal of Evaluation. *Journal of Multi-Disciplinary Evaluation, 8* (17), 57-79

Vicente, K.J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Option generation in simulated conflict scenarios in approach Air Traffic Control

Jan Kraemer[1] *& Heinz-Martin Süß*[2]
[1]*DLR Braunschweig*
[2]*Otto von Guericke University Magdeburg*
Germany

Abstract

Approach air traffic controllers provide safe guidance of aircraft approaching an airport from different arrival routes. Handling traffic and preventing separation loss between aircraft requires controllers to maintain situation awareness at all times. In case an incident is foreseen, guidance options must be acquired to deal with it. Though expert controllers are expected to always come up immediately with the best guidance options, option generation skills have often been neglected in situation awareness research so far. In addition, the fact that incidents still happen in air traffic control shows the need for research in this field. In an initial investigation study, seven expert air traffic controllers completed an online-survey consisting of videos and screenshots captured from three real-time simulations of approach scenarios on Düsseldorf airport, Germany. Every scenario was designed to end in separation loss of two aircraft. In each scenario, subjects were asked to provide as many options as possible to deal with the situation one minute prior to the incident. Results showed differences between experts regarding the quality and quantity of options successfully preventing separation loss given in the scenarios, indicating different strategies of dealing with conflict situations among subjects.

Introduction

Air travel is considered the safest mode of transportation (IATA, 2013). However, accidents still occur and with the constant growth of air traffic over the last years, the number of incidents related to air traffic management (ATM) has also increased. Statistics revealed a number of over 120 incidents per two billion flight hours in 2012 (Eurocontrol, 2013). As recent forecasts of IATA expect a total of 3.6 billion flight passengers in 2016, about 800 million more than in 2011(IATA, 2012), the number of incidents is likely to keep growing. Highly skilled air traffic controllers are needed to manage complex traffic caused by growing numbers of aircraft and to ensure safe guidance. Safety is granted by maintaining horizontal and vertical separation between aircraft within the same sector. Additionally, compliance with limitations in altitude and flight speed must be controlled at all times. Therefore, controllers constantly have to make decisions to provide safe guidance. In 2012, there have been 125 separation minima infringements per million aircraft movements (Eurocontrol, 2013). To prevent further increase, it is important to identify the sources of human error in the decision making process.

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Decision making is a cognitive process used to find the most suitable course of action (COA) among alternatives to meet a certain goal (Wang & Ruhe, 2007). Feasible COAs are derived from careful analysis of the situation dealt with. Analysing a situation's state and figuring out what to do is called Situation Awareness (SAw; Adam, 1993). More detailed, SAw has been defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1988, p. 97). Therefore, proper SAw is necessary to select the most suitable COA from a number of possible ones depending on one's objective (Endsley, 1999a). It enables an operator to anticipate the situation's future state and to direct subsequently encoding and pattern recognition accordingly (Durso, Rawson & Girotto, 2007). Knowledge about the future state has been shown to reveal the biggest differences in SAw of experts and novices (Durso et al., 1995) and is considered to be a distinct ability of skilled experts (Endsley, 2000)

Maintaining SAw while dealing with complex dynamic situations is important for good performance. Loss of SAw has been identified as the source of operational errors in air traffic control (ATC). 58.6% of operational errors in Terminal Radar Approach Control and 69.1% in enroute ATC are caused by insufficient SAw (Endsley, 1999). As SAw involves the construction of (partially) internal representations of highly complex situations, it puts effort into cognitive resources such as working memory (Baddeley & Hitch, 1974) and attention (Durso et al., 2007). Furthermore, the dynamics of moving objects require continuous updating of those representations as the situation changes over time. SAw may be lost if complexity and dynamics exceed the capabilities of the operator's attention and working memory capacity, as both are limited resources (Endsley, 1988).

Expertise can reduce the effect of limited cognitive resources on SAw (Durso & Gronlund, 1999). Subject-matter experts develop internal models from experience which help them to guide their attention, to organise information and to project future states of the situation at hand (Endsley, 1998). Those internal models are stored in long-term memory and can be activated and integrated with situation models stored in working memory (Durso et al., 2007). As they are treated as a single piece of information, they may greatly reduce the demands of storing complex information patterns (Sweller, 2003). Sohn and Doane (2004) found that expert pilots relied on their skills built from experience when recalling flight situations from given cockpit perspectives, while novice subjects' performance was determinate by their working memory capacity.

An operator needs to know his options to adequately deal with a given situation. Confronted with familiar situations, experts are believed to come up with optimal COAs from experience without having to rely on further processing (Pfaff et al., 2013). Unfamiliar situations, on the other hand, call for more complex processing if they cannot sufficiently be mapped to prior experience. According to Wang and Ruhe (2007), setting a goal triggers an exhaustive search for possible COAs and criteria distinguishing between useful decision-strategies. This is also known as decision space (DS). It results from transforming raw information from SAw to actual COAs (Drury, Pfaff, More, & Klein, 2009). By analysing the potential costs

and outcomes of available options, operators eventually decide on the most suitable one. In emergency response decision making research, assistance systems have been developed to help decision makers to find available COAs and compare them in terms of possible outcomes (Chandrasekaran, 2007). Using exploratory algorithms, such systems are designed to reduce the effects of uncertainty and time pressure in complex dynamic situations on the operator's performance. They provide the decision maker with all possible COAs, their respective outcomes and possible risks and robustness over a variety of conditions (Pfaff et al., 2013). It has been shown that such systems can improve both the accuracy and speed of identifying robust decisions from a set of alternatives (Lempert, Popper & Banks, 2003, cited by Pfaff et al., 2013).

Constructing the DS of an operator requires proper SAw and involves knowledge and expertise. SAw is affected by limited cognitive resources and must be maintained at all times to handle complex dynamic situations. Furthermore, decisions must often be made under time pressure. Given unlimited time to analyse a situation without having to memorise all the details, subject-matter experts should be able to build up sufficient SAw to deal with the situation. Thus, in combination with their expertise, they should be able to provide an enclosing set of possible COAs. In highly standardised and regulated fields such as ATC, DS are expected to bear a close resemblance among experts, because explicit rules can put an external limit to the possible options a decision maker has to find and compare. The aim of this study was to find out if experts are actually able to generate encompassing DS if they have both unlimited time and access to all relevant information. Under these conditions experts are believed not to differ in conflict resolution performance among each other. Thus, no significant differences between experts are expected in terms of both quality and quantity of their decisions.

Methods

Subjects

Ten approach and one tower air traffic controller (10 male, 1 female) from Deutsche Flugsicherung (DFS) participated in the experiment. Age ranged from 23 to 51 years (M = 32.82, SD = 8.55) while years of experience ranged from 2 to 20 years (M = 8.36, SD = 6.23). Subjects were recruited by direct advertisement via the internal network of DFS. Participation was voluntary, no expense allowance was paid.

Conflict resolution task

Subjects were asked to provide as many solutions as possible for conflict scenarios in simulated approach ATC. A computer based survey was created containing three short real-time simulated scenarios of approaching air traffic on Düsseldorf airport (EDDL), Germany. Scenarios were created using NLR ATM Research Simulator (NARSIM; ten Have, 1993), a real-time ATM simulator software developed by the National Airspace Laboratory of the Netherlands. Scenarios showed aircraft approaching Düsseldorf Airport via Standard Arrival Routes using conventional Transition To Final procedures (see Figure 1). Scenarios each lasted between four and five minutes and were designed to end in separation loss between two aircraft.

Videos of the simulation runs were recorded using desktop capturing software. Additionally, screenshots of the situation one minute prior to separation loss were captured with the respective aircraft highlighted.

Each item was introduced by a short description of the situation at hand including the time at which the conflict occurred as well as the conflicting aircraft. Underneath the introduction, the video and the screenshot of the current conflict scenario were embedded. Subjects were asked to watch the videos as well as the screenshots carefully and to find as many solutions as possible to prevent the upcoming separation loss one minute before it occurred. Separated pre-labelled tables were presented on the same page to write down advisories that would be given to the aircraft. All advisories written in one table represented one COA. Subjects were allowed to advise changes to flight speed and altitude of aircraft and to turn aircraft from the downwind to the centreline. Additionally, subjects were asked to rate if they would personally use each option in reality on a Likert scale ranging from 1 (never) to 7 (absolutely). Subjects were allowed to watch the videos and screenshots as often as they wanted and to switch back and forth between the scenarios to find as many options as possible.

Several simplifications were set in the simulation to make answers more comparable. All aircraft were set to the same type. No differences in horizontal separation had to be considered between different wake turbulence categories and no wind was present. All aircraft followed the approach procedure as described.

Procedure

The conflict resolution task was presented as an online survey. First, a biographical questionnaire was completed. Following the questionnaire, general instructions were presented, involving information about aircraft types, conventional approach procedure, how to change video settings and to fill out the direction tables. Furthermore, subjects were assured that no data could later be linked to a specific person. Thereafter subjects completed the questionnaires as described.

Data analysis

Validation of the options provided by the participants was done by replaying the scenarios once for each answer. One minute prior to the separation loss, the simulation was paused and all advisories for one solution were put into the simulation. If the separation loss was prevented successfully, the respective option was scored with one point. If any violations of limitations to speed and altitude were made, half a point was given. Zero points were given if the conflict still occurred or new conflicts were produced. Options were categorized by combinations of directions given. This way, small deviations in absolute values assigned between subjects did not count as distinct options.

Results

Subjects provided a total of 12 original options in total for scenario one. Ten options were found in scenario two and seven in scenario three. Descriptive statistics of

valid options provided per subject in each scenario are presented in Table 1. No significant deviations from either uniform or normal distribution were found using Kolmogorov-Smirnov-Tests in any scenario. Paired t-tests showed significant deviations of mean numbers of valid options per subject from total valid options provided by all subjects in scenario one ($t(10) = -28.03$, $p < .001$), two ($t(10) = -33.80$, $p < .001$) and three ($t(9) = -21.10$, $p < .001$).

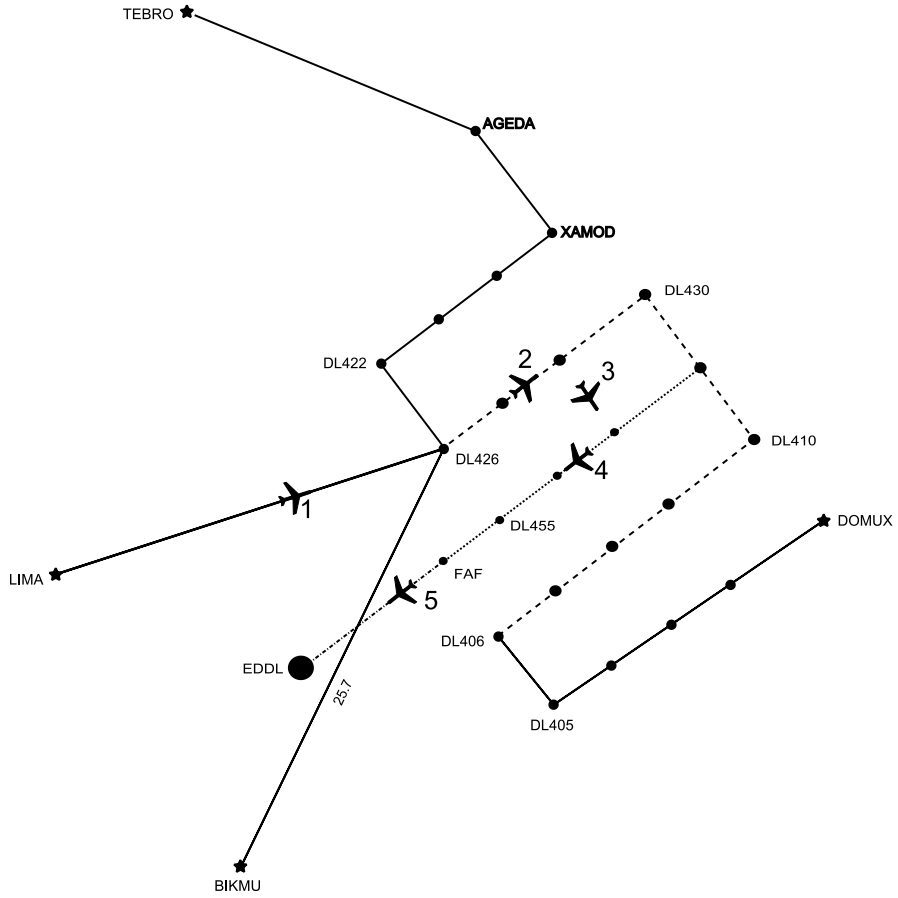


Figure 1. Schematic depiction of conventional Transition To Final procedure at Düsseldorf Airport (EDDL), Germany

In order to take differences in experience into account, subjects were divided into two groups by median split ($Mdn = 6$). A 3x2 ANOVA with scenarios as within-subject factor and experience (low vs. high) as between-subject factor revealed significant differences in the total number of options provided per subject among scenarios ($F(2, 18) = 7.43$, $p = .005$, $\eta^2 = .48$). Post-hoc Bonferroni-corrected paired t-tests showed no significant differences between scenarios. No significant effects of scenarios ($F(2, 18) = 2.87$, $p = .086$) or experience ($F < 1$) on mean numbers of valid

options were found. No significant correlation between experience and valid options were found throughout scenarios (r_s = .22, p = .257*).*

Table 1. Distribution of valid options provided per scenario (Top – Total number of valid options provided over all subjects, % valid – percentage of valid answers given, KS-Z Kolmogorov-Smirnov-Z values of tests for uniform distribution)

Scenario	Top	*MD*	*SD*	% valid	Min.	Max.	*KS-Z*	*p*
1	12	2.0	1.2	70.9	0.0	4.0	0.53	.943
2	10	1.2	0.8	46.2	0.0	2.5	0.63	.819
3	7	1.5	0.8	72.5	0.0	3.0	0.74	.648

Frequencies of common options provided by subjects were counted for each scenario (see Figure 1). Out of the 12 options in scenario one, four options were stated by more than two participants. One option out of ten in scenario two and three out of seven options in scenario three were used more than twice.

No significant correlation between ratings and validity of options were found among scenarios ($r(75)$ = -.05, p = .348). Paired t-tests between mean ratings of valid and invalid options showed significantly higher ratings of valid options in scenario two ($t(5)$ = 4.72, p = .005, d = 1.42). No significant differences of mean ratings were found in scenario one ($t(2)$ = -0.28, p = .808, d = -0.17) and three ($t(3)$ = -2.85, p = .065, d = -1.88) Out of the eight common options used by three or more subjects, ratings of four solutions differed no more than two points. Ranges in the remaining options went up to a maximum of five points.

Discussion

It was expected that various subject-matter experts would provide highly similar sets of possible COAs when confronted with the same conflict scenarios. Data showed that this is not the case, even though Kolmogorov-Smirnov-Tests showed no significant differences in quantity of given and valid answers. As no deviation from standard distribution was found as well, the results indicate that the tests lacked significance due to the low power arising from the small sample. Various experts came up with a lot of different solutions to the same situation. Furthermore, none were even close to providing all possible solutions in any scenario. While a total of 12, ten and seven different options were given in total, the maximum number of experts sharing one option was never higher than four among scenarios. Moreover, some high differences between ratings for the same options were found among subjects. This is surprising considering that all of the participants were highly trained professional air traffic controllers. Insufficient SAw as a cause of error was unlikely as unlimited time was given to solve each scenario and all relevant information was accessible throughout the task. Additionally, no information had to be memorized over longer periods of time because videos and screenshots could be watched repeatedly. Nevertheless, subjects not only failed to provide complete DS but even produced invalid solutions which did not solve the respective conflicts.

As subjects' experience covered a range of 18 years, this might explain differences in DS. As experience and knowledge were discussed as important factors in acquiring SAw, an increasing number of correct answers would be expected with higher experience. However, the data do not support this explanation as no higher scores were found for the more experienced subjects. Nevertheless, it should be kept in mind that this might stem from the small sample, namely the lack of statistical power as mentioned. In a larger sample, experience might make a difference when it comes to finding solutions in emergency situations.

The available advisories and simplifications used during the experiment may pinpoint another explanation for the differences in experts' DS. Some subjects criticised the lack of heading advisories claiming that this eliminated possible options. In that case, experts should have been even more likely to produce similar sets of COAs due to the reduced amount of options left. The low level of compliance found among the answers provided throughout the experiment contradicts this point. Although options were excluded from the start, subjects still came up with a lot of different approaches to the same problems and differed strongly in both quantity and quality of their answers. Allowing for more directions might have resulted in even bigger variance of both. Unfortunately, it was not possible to test this supposition with the acquired data.

The low number of options may result from subjects tending to provide only robust COAs instead of encompassing DS. It has been argued that optimal COAs are almost impossible to find in complex dynamic situations due to their high levels of uncertainty and time pressure (Lempert et al., 2003, cited by Pfaff et al., 2013). Therefore, decision makers tend to make robust decisions which maintain their effectiveness over a wider range of possible outcomes and conditions in emergency situations (Bryant & Lempert, 2010). However, two findings in this experiment contradict this explanation. First, subjects were only watching a simulation and were given as much time as they wanted to produce their answers. Therefore, it is unlikely that time pressure kept them from thinking all their options through or rushed them towards making decisions. Second, subjects also provided options rated with only two points (very unlikely), meaning they gave an answer they would not really use in a real-world situation.

Subjects may have provided fewer answers than they could possibly have due to lack of motivation. As the task required them to rethink a situation over and over to come up with new ideas, this might have reduced compliance with the task over time even though participation was voluntary. Indeed, the descending number of total options provided per subject among scenarios indicates loss of motivation throughout the task. On the contrary, no decrease in valid options was found between scenarios. Loss of motivation may explain why fewer answers were provided in the last scenario. However, it does not explain why the quality of the answers did not drop over scenarios. Due to anonymity, contacting participants in order to confront them with the results and ask about problems afterwards was impossible. Future studies of this kind could be combined with post experimental interviews to allow for more detailed explanations of strategies used to identify

possible COAs. Additionally, allowing subjects to further explain their answers might help to keep up motivation throughout the task.

It has been argued that current air traffic systems will not be able to cope with projected increases in air traffic due to lack of flexibility (Lohr & Williams, 2008, cited by Pfaff et al., 2013). Assistance systems which have been developed and are already in use by some air navigation service providers may help air traffic controllers to overcome this problem by providing a broader range of COAs (Pfaff et al., 2013). Looking at the data, the question arises if such systems should already be mandatory for emergency decision making in ATC. Although unlimited time was given to solve each scenario, subjects still produced invalid answers which didn't prevent the conflicts. In addition, in each scenario at least one subject failed to produce any valid options at all. In future studies, it would be interesting to compare the DS of human experts directly to emergency assistance systems which make use of robust decision making processes. If all possible COAs and their estimated outputs were derived from modelling processes, it could be tested if human experts are able to provide a similar set of answers. Additionally, it could be examined if DS of human experts, although they may be smaller in quantity, are representing the most robust COAs found by the assistance systems. Unfortunately, such systems were not available in this study. Furthermore, although it has been argued that lack of SAw was an unlikely cause of error in this study, this cannot be ruled out. Future work should include the assessment of SAw data using probe methods such as the Situation Present Assessment Method (Durso, Blackley & Dattel, 2006) to draw more resilient conclusions about Saw and DS generation.

References

Baddeley, A.D., & Hitch, G.J. (1974): Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York: Academic Press.

Bryant, B., & Lempert, R.J. (2010). Thinking Inside the Box: A Participatory, Computer Assisted Approach to Scenario Discovery. *Technology Forecasting and Social Change, 77*(1), 34-49.

Chandrasekaran, B. (2007). *From Optimal to Robust COAs: Challenges in Providing Integrated Decision Support for Simulation-Based COA Planning*. Laboratory for AI Research, The Ohio State University.

Drury, J.L., Klein, G.L., Pfaff, M., & More, L. (2009). Dynamic decision support for emergency responders. In *Proceedings of the 2009 IEEE Technologies for Homeland Security Conference* (pp. 537 – 544). Waltham, USA: Technologies for Homeland Security.

Durso, F.D., Blackley, M.K., & Dattel, A.R. (2006). Does Situation Awareness Add to the Validity of Cognitive Tests? *Human Factors, 48*, 721 – 733.

Durso, F.D., & Gronlund, S.D. (1999). Situation Awareness. In F.T.R.S. Nickerson, S.T. Dumais, S. Lewandowsky, and T.J. Perfect. (Eds.), *Handbook of Applied Cognition* (2nd ed., pp. 163-193). Chichester, UK: Wiley.

Durso, F.T., Truitt, T.R., Hackworth, C., Crutchfield, J., Nikolic, D., Moertl, P (1995). Expertise and chess: A pilot study comparing situation awareness methodologies. In D.J. Garland & M.R. Endsley (Eds.), *Experimental analysis and measurement of situation awareness* (pp. 295–304). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Endsley, M. (1988). Design and Evaluation for Situation Awareness Enhancement. In *Proceedings of the Human Factors and Ergonomics Society 32nd Annual Meeting* (Vol. 1, pp. 97 – 101). Anaheim, CA: Human Factors Society.

Endsley, M. (1999). Situation Awareness and Human Error: Designing to Support Human Performance. In *Proceedings of the High Consequence Systems Surety Conference*. Albuquerque, NM: Sandia National Laboratory.

Endsley, M. (1999a). Situation Awareness In Aviation Systems. In D.J. Garland, J.A. Wise, and V.D. Hopkin, (Eds.), *Handbook of Aviation Human Factors* (pp. 257-276). Mahwah, NJ: Lawrence Erlbaum Associates.

Eurocontrol (2013). *Annual Safety Report 2013.* Retrieved from http://www.eurocontrol.int/sites/default/files/content/documents/single-sky/src/src-docs/src-doc-53-e1.0.pdf

Gheisari, M., & Irizarry, J. (2011). Investigating Facility Managers' Decision Making Process through a Situation Awareness Approach. *International Journal of Facility Management, 2*, 1-11.

International Air Transport Association (2013). *IATA Annual Review 2013*. Retrieved from http://www.iata.org/about/Documents/iata-annual-review-2013-en.pdf

Pfaff, M.S., Klein, G.L., Drury, J.L., Moon, S.P., Liu, Y., & Entezari, S. (2013). Supporting Complex Decision Making Through Option Awareness. *Journal of Cognitive Engineering and Decision Making, 7*, 123-140.

Sohn, Y.W., & Doane, S.M. (2004). Memory Processes of Flight Situation Awareness: Interactive Roles of Working Memory Capacity, Long-Term Working Memory, and Expertise. *Human Factors, 46*, 461-475.

Sweller, J. (2003). Evolution of Human Cognitive Architecture. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 43, pp. 215-266). San Diego, CA: Academic Press.

Wang, Y., & Ruhe, G. (2007). The Cognitive Process of Decision Making. *International Journal of Cognitive Informatics and Natural Intelligence, 1*, 73-85.

The operational potential of an In-Flight Weather Awareness System: an explorative pilot-in-the-loop simulation

Simone Rozzi[1], Stefano Bonelli[1], Ana Ferreira[1], Linda Napoletano[1], & Loic Bécouarn[2]
[1]Deep Blue s.r.l., Rome, Italy
[2]Thales Avionics, Bordeaux, France

This study investigates the operational potential of an in-flight weather awareness system displaying weather hazard cues that are either invisible (i.e. Clear Air Turbulence and Icing) or visible only during clear visibility operation (i.e. Cumulonimbi, and Volcanic Ash). The study focuses on investigating (i) the potential uses of the display, (ii) its usability deficiencies, and (iii) its potential for pilot error. Methodology: A small-scale human-in-the-loop simulation coupled with expert observations, followed by a questionnaire and in-depth interviews. A total of 14 professional pilots flew several scenarios using the evaluated display to plan route changes free of weather conflict. Results: The display exhibits the potential to shift weather management from a tactical (5–10 minutes) to a strategic level (up to 1h earlier than today). Cluttering due to multiple overlapping weather areas was the main usability deficiency. Mode error could occur due to poor indication of weather hazard status, and when using the proposed display in less modern airspaces than Europe and US. Value: These findings are relevant for human factors and safety specialists and researchers involved in the development, evaluation, purchase and certification of aviation weather displays.

Introduction

For operators of complex systems it is important to respond effectively to the hazardous events that can affect the safety and efficiency of the processes they control. In aviation, the availability of digital displays offers a unique opportunity for safer and more efficient pilot's response to weather hazards. At the same time the development and introduction of any of such displays calls for a thorough evaluation of their actual impact in the context of use, i.e. the flight deck.

This paper presents a small-scale pilot-in-the-loop simulation aimed at evaluating the operational potential of an in-flight weather awareness system. This system provides pilots with a large-screen and intuitive view of the flight 4D trajectory—the three spatial dimensions of aircraft trajectory plus time—complete with the surrounding weather hazards.

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Background

Weather and aviation

Weather is a long standing source of disruption in aviation. Besides causing delays, excessive fuel costs and lost passenger time, weather continues to be an important safety concern. NTSB statistics see it as a primary contributory condition in the 23% of aviation accidents (Kulesa, 2003). Adding to this, weather-related accidents are far more likely to cause fatalities than accidents that occur in visual meteorological conditions (NTSB, 2005).

Important hazardous weather events include encounters with (i) cumulunimbi clouds, (ii) clear air turbulence, (iii) icing and (iv) volcanic ash. Cumululninbi clouds (CB) can cause excessive turbulences, can interfere with communication and navigation systems, and can even lead to engine failure. The consequences of an encounter with clear air turbolences (CAT) can vary from slight discomfort for passengers to potential for structural damage, impaired crew performance and injuries for passengers and cabin crew members (Airbus, n.d.; SKYbrary, 2014). In-flight icing (ICE) occurs when ice accumulates on exposed and unprotected surfaces of the aircraft: this effect can disrupt the smooth flow of air over the wing, thus degrading lift; can generate false instrument readings; and can also compromise the handling qualities of the aircraft. Encounters with vulcanic ashes (VA) can result in engine damage and malfunction, since particles can melt within the engine or even disturb the airflow.

When encountering these weather events along the course of the flight, pilots have to devise diversions from the planned flight plan to circumnavigate these events while ensuring adequate separations from them. One crucial aid to support this reasoning is the on-board weather radar. However, one important limiting factor with this system is the shadowing effect: radar waves are reflected by droplets, so when facing a CB it is not possible to see what is behind it—radar waves are blocked by it. As a result pilots might may change the flight path in a way that can turn out to be inadequate the moment they realize what is behind the CB line (Craig, 2012). Also non-technological weather information sources include information provided by Air Traffic Control (ATC), which can inform pilot of Pilots In-Flight Reports (PIREPs) broadcasted by aircraft that have passed previously in the same area of interest. Unfortunately, this information is based on subjective judgement. Also, the pilot gain information about the weather picture during the initial mission planning phase of the flight. However, weather may evolve since the start of the flight.

ALICIA WAS

To address the above limitations on in-flight weather management, the ALICIA project (*All Conditions Operations and Innovative Cockpit Infrastructure*), an EU cofounded project in the FP7, has proposed a novel Weather Awareness System (WAS) that displays information about the athmospheric hazards along the 4D trajectory of the flight. The display is touch enabled and is composed of two views (see also Fig. 1):

- *The top view:* this is the larger view and provides pilots with a 2D birdeye picture of the current flight path and the surrounding weather situation (see also Fig. 2). It uses a different symbology and colour coding for each weather events: CB are displayed as yellow areas; CAT as magenta small arrows; ICE as blue areas; and VA as dark-to-grey areas depending on ash concentration—with dark representing the higher VA concentration and most dangerous zones.
- *The lateral view:* located below the top 2D view, this view portrays the vertical profile of the flightpath togheter with the weather events that will cross this path. Colour coding and symbology are the same as for the top view, except that this view shows also the vertical extension of the weather events.

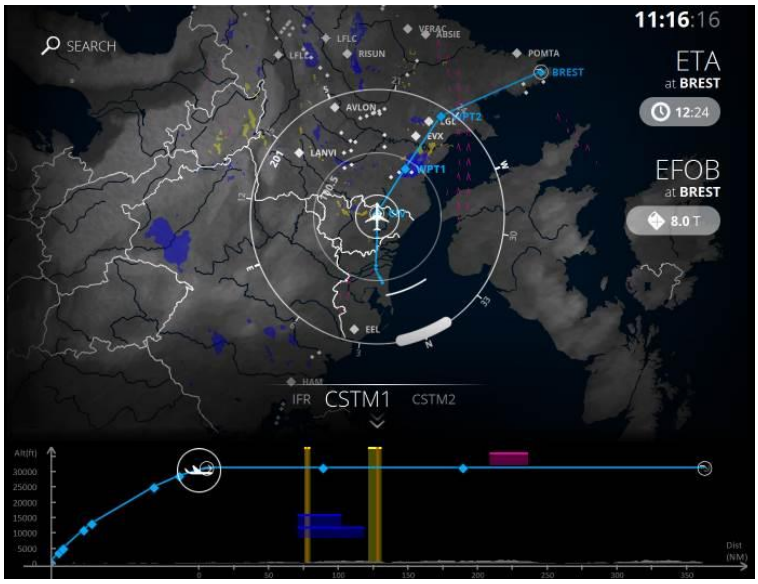


Figure 1. Top and lateral views of the evaluated display.

On both displays, weather events visualization is not fixed. To avoid cluttering, pilots can choose which weather hazard to visualize by pressing the corresponding touch screen button available on a dedicated menu. When activated, this menu appears over the lateral view. The system is based on ground meteorological data uplinked to the aircraft and it displays the current weather situation (nowcast). The future weather situation (forecast) can be displayed acting on a time-line provided on the right part of the display (not working during the study). The system automatically calculates future conflicts along the flightplan and displays them as red triangles placed on the expected conflict point. It checks the forecasted weather situation along the route according to the future aircraft position (based on the flight management system). Touching the conflict point a menu can be opened and a new route can be calculated by the system and showed as a tick white line. This new route is conflict free and it returns to the original flight plan as soon as possible.

Pilots can accept the new route (and send it to ATC via datalink to be cleared) or refuse it.

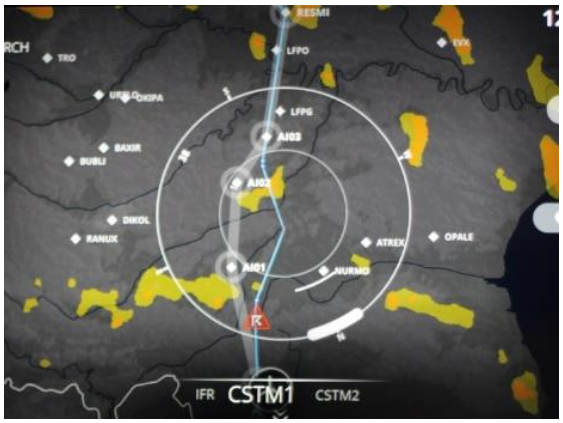


Figure 2. A close up view of the ALICIA WAS top view.

Theoretical perspective

The actual benefit that the proposed ALICIA display can deliver requires a qualitative understanding of how this system interacts with the intended operational context. Notably, a new interface technology is introduced under the lure of quantitative and measurable performance improvements in areas such as capacity, cost-effectiveness, and safety. However, the ability of the technology to deliver on these areas requires consideration of how the new system can be actually used by the human practitioner in context. Two arguments can be made in support of this point. First, stakeholders located at higher hierarchical level than operations, such as senior management, modernization and programme leaders, may tend to assume a direct linear relationship between the new technology and the desired improvements, i.e. the provision of the new technology will attain the desired system (wide) objective(s).

However, such an assumed relationship may actually prove simplistic when compared against the reality of the operational context in which new technology will be used. Technology is imperfect, and operational experience indicates it can be clumsy and difficult to use. It is often the case that the practicalities of fitting the new technology to the work needs to be worked out by practitioners themselves in order to adapt it to the work environment and get the job done (Cook, Nemeth, & Dekker, 2008; Cordesman & Wagner, 1999; Demchak, 1991). One of the possible outcomes of this tailoring process is the use of the new technology for purpose(s) other than the intended prescribed one(s) (e.g. using the on-board anti collision alert system as separation aid). Furthermore, technology can change the nature of the work in undesirable ways: it can redistribute workload unevenly across different phases of the flight, it can deskill human operators, it can make the functioning of the system more obscure to its user, it can increase the potential for surprise (e.g. Chialastri, 2012; Degani, 2004; Sarter, Woods, & Billings, 1997; Bainbridge, 1983).

In safety critical domains, changes such as these may introduce new paths to disastrous failure that did not exist prior to the introduction of the new technology (Strauch, 2004; Woods, Dekker, Cook, Johannesen, & Sarter, 2010).

These considerations lead to the second argument: the actual use of new technology is not something that can be easily assumed or anticipated without appreciating the situational or contextual perspective of the end user. New automated systems are not introduced in a vacuum in fact, but into an existing on-going field of practice made up of people and technological artefacs (Woods et al., 2010). Here, the human strives to meet the multiple and often conflicting demands of the job under intense organizational pressures for productivity, high environmental uncertainty, and limited attentional and temporal resources (Hollnagel, 2012; Hollnagel & Woods, 2005). Thus, the potential of new technology requires consideration of the expert and contextual view of human operators: because they have a first-hand direct understanding of their field of practice, of its intrinsic complexities, trade offs, demands, and uncertainties (Dekker, 2004), operators are best placed to know how they will use the new artefact, for which purposes and which problems may arise in the process. These aspects are not easily intelligible for stakeholders located at higher organizational levels, such as management and engineering, as they lack temporal and spatial proximity to the complex dynamics of the operational environment.

Objectives

The above considerations emphasize the importance of conducting qualitative explorations about the potential of new technology in a way that accounts for the view point of the expert practitioners (the target user) since the very early developmental stage. This is particularly important in the case of technology-centered development processes (Boy, 2012), which may lack a torough exploration of the role of the novel technology prior deployement. The present study aims at exploring the interaction between the proposed display and the target operational context. In particular it focuses on investigating**:**

(i) The potential uses of the display, i.e. what pilots believe they could do with the system;
(ii) Its usability deficiencies, i.e. which aspects of the display may hamper access and manipulation of information;
(iii) Its potential for human error, i.e. what error can occur during the use of the display.

Methodology

Location and Equipment

This study was conducted at Thales Avionics over the period Sept 13–Jan 14 in Bordeaux. It made use of a two-person crew fixed based cockpit simulator called Avionics 2020. The evaluated display was located on a central head down

navigation display (size=19inch) that was visible to both the non-flying and the flying pilot, as shown by Fig. 3.

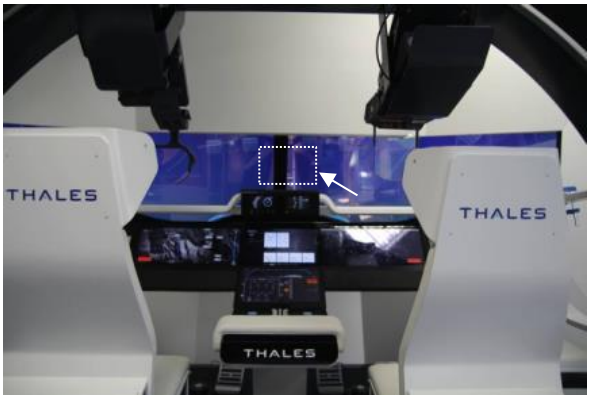


Figure 3. The position of the ALICIA WAS in the simulator used for this study.

Participants

14 professional pilots from three European airlines and two European aircraft manufacturers participated voluntarily in the evaluation. Three pilots had previous military experience as jet fighter pilots. Flying experience ranged from a minimum of 2600 flight hours to a maximum of 20000 flight hours, with an average of 8960 flight hours. All pilots were men, their average age was 53 years, wth the oldest participant being 68 years old and the youngest 35 years old (sd=10 years). All pilots were familiar with electronic displays. All but two pilots were familiar with touch screen. All but four pilots reported to have flown with head up displays. The participants provided their written consent to participate in the study, and completed a biographical questionnaire.

Scenarios and Task

Three En Route scenarios were played: a flight from Amsterdam Schiphol to Clermont Ferrand Auvergne with CB encounters; a flight from Amsterdam Schiphol bound to Brest Britagne airport with CAT and ICE encounters; a flight from Barcelona to Istanbul with VA encounter. Each scenario lasted approximately 30 minutes and was flew by a crew of two. At the start of the scenario the crew was requested to use the ALICIA display to devise collaboratively potential route changes to their planned flight plan. The crews were also invited to explore the various display functionalities, and to report outloud their opinions and criticisms about the value of the displayed information features and the quality of interface management.

Data collection

During each simulation run, human factors researchers took observational notes of pilots' behaviours. These captured the unfolding pilot interaction with the display, pilot-to-pilot interactions, as well as pilots' comments and impressions about the evaluated system.

After completing the three runs, the pilots completed a questionnaire. This collected biographical data and ratings to ten items that evaluated pilots' perspective on these areas: safety, situation awareness, weather conflict avoidance, punctuality, efficiency, workload, usability, basic task, standardization. Each rating was on a 5-point-Likert scale (1=highly disagree; 5=highly agree). The questionnaire was refined before applying it in the study and was administered on line: this means that participant ratings were available for subsequent interviews and the final de-briefing.

Upon completion of the questionnaire, the rationale behind each rated item was probed by means of in-depth interviews. These developed consistently with the principles of the Critical Decision Method (Hoffman, Crandall, & Shadbolt, 1998; Klein, Calderwood, & Macgregor, 1989): whenever a pilot reported a display problem or benefit, he was prompted to think of a relevant real life scenario to explain what role the display could play, considering the specific scenario demands, constraints, available information cues usually attended, and the likely mistakes that coud occur if things go wrong. During this process, pilots were invited to sketch the described situation to clarify the the underlying spatial-temporal reasoning. Each interview lasted between 30 to 45 minutes. They were recorded and transcribed.

Data Analysis

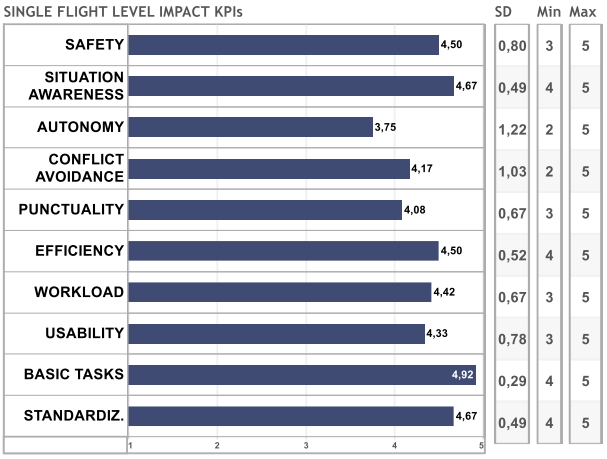
Descriptive statistics was used to report the questionnaire results. The Emergent Theme Analysis (Wong & Blandford, 2002) approach was used to analyse interview data. This qualitative method is suitable for making sense of large interview data about expert knowledge in safety critical domains. Initially the data was searched for broader themes, i.e. meaningful portion of the data that in this study captured capabilities and limitations of the evaluated display. Subsequently, the data was searched for sub-themes, i.e. data fragments that support and allow to refine the higher level broader theme they belong to. Sub-themes indentification and descriprion made use of a framework composed by four categories: aircraft situation, demand for the pilot, available information cues, and role of the display in the specific situation. After completing the analysis, early results have been presented to the participating pilots for corroboratory purposes during a one-day post-simulation meeting.

Results

Questionnaire ratings in Table 1 indicate that the participating pilots assigned high scores to almost all of the investigated aspects (agreements rates are between 4 and 5 for all statements). In particular, the areas of safety, situation awareness, efficiency and workload are rated quite high, thus indicating that the display was perceived to bring a positive impact to the management of weather. Autonomy was cautiously

agreed as pilots regard weather related decisions as a pilot's decisions that can be done without ATC support—and the ALICIA WAS does not alter this situation. The area of basic tasks received the highest rating and refers to the fact the the display was viewed as not disruptive of existing cockpit activities.

Table 1. Questionnaire category mean, standard deviation, and minimum and maximum. Likert Scale (1=highly disagree; 5=highly agree).



Potential uses

The analysis of the interview data was instrumental to interpret the questionnaire ratings. Pilots reported that, compared to today systems, ALICIA WAS can help them to build a comprehensive and intuitive long range picture of the current weather situation, from departure to arrival, that is directly functional to CB, ICE, CAT and VA identification and avoidance. In particular, the following uses have emerged from the study.

C1. Formulating a global diversion, instead of a small range one
Whenever possible, pilots are interested on devising an alternate route clear of conflict from all of the weather hazards that may exist along the originally planned route—rather than implementing minor short range changes to this latter. This latter strategy lacks cost effectiveness because it exposes pilots to the risk of implementing a short range but ineffective change, which requires further close monitoring and adjustement. ALICIA WAS was reported to support the demand for formulating a global diversion because it allows pilots to see the complete weather picture, from departure to arrival. This is information is not available with the current on-board radar.

C2. Anticipating weather management
Pilots commented that the long range weather picture provided by ALICIA WAS facilitates pilot assessment (i) of the existence of dangerous weather conditions at longer distances, and (ii) of the level of threat these pose to current flight route. In turn pilots can make more strategic decisions concerning what should be done—i.e.

formulating a diversion versus continuing the flight without changes—much earlier compared to today's operations. In their view this was the most significant advantage offered by the display. When asked about how much earlier they could start considering the best (alternative) path considering the current weather situation, pilots reported that that the ALICIA WAS could allow them to think about weather related diversions from 30 minutes to 1h in advance compared to today.

The cost of anticipating weather related decisions is that more effort will be spent by the crew for identifiying the best route when still relatively far from adverse weather areas; however, pilots reported that this addedd effort is desirable because it can drastically reduce the risk of entering an adverse weather area. This latter is an undesirable situation that places a high burden on pilots to restore the safety of flight.

Also, pilots reported that the potential for anticipating weather management can be greatly enhanced by complementing the current version of the display with information about weather (i) historical evolution and (ii) future evolution. Especially for cumulunimbi, to pilots it is important to understand the growing or expansion rate of these events on both the vertical and the horizontal dimensions. This is particularly important when flying over tropical areas, for there weather fronts can grow very quickly in a short amount of time. Depicting past and future information about the evolution of large (and highly dynamic) CB increases pilot ability to formulate a single successful lateral diversion, i.e. it decreases the risk of selecting a diversion that although appropriate at present time, considering current CB dimensions, will need to be modified at a later time as it intersects the expanded volume of the same CB, which has grown wider in the meantime;

C3. Identifying the best airport to descend to in case of emergency.
Three pilots reported that the system could be helpful during emergency situations to evaluate the weather conditions close to the ground. The display would facilitate and support the choice of the best airport where to land in case of an emergency, considering current position, weather situations, underlying terrains and aircraft (decreased) capabilities. Also the system could be useful during engine out situations, especially when flying over high terrains, to check readily whether there are cumulunimbi or other weather hazards at the maximum flight level that can be sustained by the aircraft.

Missing information cues

Pilots noted that the system was not ready for operational use. They suggested a range of missing information items that need to be provided so that they can work with the system. These are listed below:

- *Weather Information age.* To trust and use the system pilots need to know how old the displayed information is, i.e. when it was calculated. They reported to be afraid of making decisions about diversions based on information that is not valid anymore by the time the decision is made;
- *Width of the section of airspace displayed on the vertical display.* A further missing information was the width of the section of airspace represented in the

vertical display (see Fig. 1). As this display is 2D, depth is not represented, thus pilots miss essential information cues both about (i) the horizontal distance between the displayed weather events and the trajectory of the flight, and (ii) the width of these events.

- *Contextual Weather Information.* Pilots noted that it would be useful if ALICIA WAS supported further inspection of the weather conflict points identified and displayed by the system along the 4D trajectory of the aircraft. Currently, the existence of a weather conflict is signalled by the "R"icon (see Fig. 2). Clicking on each displayed conflict point the pilot can see pieces of information such as time and distance to conflict. Additional details could be provided—such as altitude and severity level of the hazardous weather event in question—so to make the display more informative;

Usability

Cluttering induced by colour coding deficiencies was the main usability problem. It occurs when multiple weather areas, i.e. CB, ICE, Turbulence, overlap on the same area of the display. Pilots suggested implementing a filtering function that allows selecting weather events only within a given range, e.g. 1000 feet below and above a given flight level. Aggravating the cluttering problem were the borders of the countries depicted on the map.Their thickness made them unnecessary salient for pilots. Besides cluttering, pilots raised a set of colour coding issues: they favoured a representation of CB areas complete with marked CB boundaries, as this is more consistent with their visual experience of CB as seen from the cockpit seat. Also, they required a more salient colour for ICE, as they would not normally associate blue with a threat.

Potential for Error

E1: Error Mode: Pilot forgetting the weather visualization when set as idle
Pilots might fail to notice a CB, CAT, ICE, or VA because s/he might forget that the visualization of any of these weather events has been set as idle. Two conditions of current HMI design can lead to such error: first, no information cue about the visualization state (on/off) of weather events is displayed on the horizontal (strategic) top view of the ALICIA WAS. At the same time this is the area where pilot's attention concentrates the most in order to acquire weather information. Second, the control panel grouping the touch screen buttons enabling to switch on/off weather events visualization is hidden below the vertical display and is not normally visible if not intently selected. Thus, these two conditions might result in pilots losing track of the selected HMI setting, consequently failing to realize that a relevant weather hazard is not visible only because he or she has not activated its visualization. Partially mitigating this apects is the fact that even if the display of weather objects is not selected, the system will raise an alert if there is an expected conflict with the flight plan;

E2: Over Trust: relying on the system when flying over underequipped airspaces

The continual use of a reliable ALICIA WAS might lead pilots to get used to trusting this system also when flying over regions not enabled with the necessary ground based infrastructure. Pilots envisaged this situation could occur for instance when a pilot normally flying in Europe or US flies over some less equipped regions in Africa or the Middle East. As no weather information would be supplied from the ground to the ALICIA WAS, the crew might think that no weather hazard ahead exist when in fact it does—and s/he could actually be flying into it.

Discussion and conclusion

This study has explored the operational potential of an in-flight weather system by means of a small scale pilot-in-the-loop simulation. The study has provided "a preview" into how pilots' activities may change following the introduction of the evaluated display. In particular, the system was reported to provide pilots with an intuitive long range global view of the weather situation encountered by the aircraft. This can allow pilots to formulate a global diversion when facing hazardous weather events, instead of a short range one. In particular in tropical areas, the display can protect the aircraft from the risk of missing a farther and larger weather front that is rapidly growing behind the closer and smaller CB in front of the aircraft. This can reduce the risk for the aircraft to fly unintentionally into a larger (hidden) storm after having avoided a first CB. Also, the display was reported to have the potential to anticipate weather managements to 30 minutes–1h compared to today, thus shifting weather management from a tactical to a strategic level. Finally, during emergencies, the display can be helpful to select the airport whose weather conditions are more favourable for an emergency landing. These capabilities are directly relevant to the management of weather hazards as today they are not supported by the existing on- board radar.

On the negative side, the display was not considered mature for operational use, for it lacks fundamental information such as weather information age, width of the section of airspace represented in the vertical displays, and contextual weather information. Cluttering due to poor colour coding was the main reported usability problem. Errors in the use of the display could occur if the pilots forgets to turn on weather event visualization, and in case the aircraft flies into sub-equipped airspaces, as these may lack the ground weather data required by the ALICIA WAS.

Overall, these findings provide information useful for evolving the evaluated display concept further—i.e. up to a maturity level appropriate for operational testing and subsequent certification approval. One important aspect to consider is the reliability of the ground based data: although this aspect was assumed to be satisfactory for the purpose of the present study, it will need to be addressed by future developments and evaluations. Beyond the context of this study, the identified findings can provide an initial benchmark available to practitioners involved in the development, deployement and monitoring of weather displays.

From a methodological perspective, the study has the merit of having illustrated one viable approach to explore the operational role of a low maturity display concept.

Literature on safety critical automation suggests that it is important to understand (i) how a novel technology can be used in the field of practice and for which purposes, and (ii) what qualitative changes it can bring. However, the introduction of novel safety critical technologies may neglect the consideration of these aspects; as a result the new technology may be used in ways that deviate from the envisaged and prescribed use, and may introduce new paths to failure. These problems occurs because the development of novel safety critical displays is usually technology centred (Boy, 2012; Jackson, Dorbes, & Pinchacourt, 2000): it is driven by the availability of new technological capabilities, so that engineering development precedes the detailed search of actual potential uses. For instance, in the present case, display development was mostly propelled by the availability of novel enabling communication and data base technologies. In contexts such as these, the use of qualitative human-in-the-loop simulation emphasizing the in-depth understanding of the perspective of expert practitioners seems a plausible approach to shed light into how the new system can fit into the field of practice—in term of its potential uses, usability problems and errors. Such understanding, which has to be refined throughout system lifecycle as new issues emerge, can arguably help leaders and professionals involved in the development, deployment and management of the novel technology to develop more realist expectations about what potential the new system can deliver.

Acknowledgement

The authors wish to acknowledge the participating pilots, the Thales engineers and personell involved in the preparation of the simulation platform, and Wim Huson, from Use2aces, for his support during the set up of the simulator, and the evaluation and refinement of the questionnaire and interview protocols.

References

Airbus. (n.d.). Turbolence Threat Awareness. Airbus. Retrieved from http://www.airbus.com/fileadmin/media_gallery/files/safety_library_items/AirbusSafetyLib_-FLT_OPS-CAB_OPS-SEQ10.pdf

Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*, 775–779.

Boy, G. (2012). *Orchestrating Human-Centered Design*. London: Springer.

Chialastri, A. (2012). Automation in Aviation. In K. Florian (Ed.), *Automation* (pp. 79–102). Rijeka: InTech.

Cook, R., Nemeth, C., & Dekker, S. (2008). What went wrong at the Beatson Oncology Centre. In E. Hollnagel, C. Nemeth, and S. Dekker (Eds.), *Resilience engineering perspectives. Vol. 1: Remaining Sensitive to the Possibility of Failure* (pp. 225–236). Farnham: Ashgate.

Cordesman, A.H., & Wagner, A. (1999). *The Lessons of Modern War, Volume IV: The Gulf War*. Boulder, CO: Westview Press.

Craig, C. (2012). Improving flight condition situational awareness through Human Centered Design. *Work: A Journal of Prevention, Assessment and Rehabilitation*, *41*, 4523–4531.

Degani, A. (2004). *Taming HAL: Designing Interfaces Beyond 2001*. New York: Palgrave Macmillan.

Dekker, S. (2004). *Ten Questions about Human Error: A New View of Human Factors and System Safety*. Hillsdale NJ: Lawrence Erlbaum.

Demchak, C.C. (1991). *Military Organizations, Complex Machines: Modernization in the U.S. Armed Services*. Ithaca, N.Y.: Cornell University Press.

Hoffman, R.R., Crandall, B., & Shadbolt, N. (1998). Use of the Critical Decision Method to Elicit Expert Knowledge: A Case Study in the Methodology of Cognitive Task Analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(2), 254–276.

Hollnagel, E. (2012). *The ETTO Principle: Efficiency-Thoroughness Trade-Off*. Farnham: Ashgate.

Hollnagel, E., & Woods, D.D. (2005). *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. New York: CRC Press.

Jackson, A., Dorbes, A., & Pinchacourt, I. (2000). Striving for Adequacy: The Importance of Rich HMI Requirements. Presented at the 3rd USA/Europe Air Traffic Management R&D Seminar, Napoli, Italy.

Klein, G.A., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. *Systems, Man and Cybernetics, IEEE Transactions on*, *19*, 462–472.

Kulesa, G. (2003). Weather and aviation: How does weather affect the safety and operations of airports and aviation, and how does FAA work to manage weather-related effects? Presented at the The Potential Impacts of Climate Change on Transportation Workshop, Washington DC.

NTSB. (2005). *Risk Factors Associated with Weather-Related General Aviation Accidents.* (No. SS0501). Washington, D.C.: National Transportation Safety Board.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of Human Factors and Ergonomics*, *2*, 1926–1943.

SKYbrary. (2014). Clear Air Turbolence [Electronic repository]. Retrieved May 9, 2014, from http://www.skybrary.aero/index.php/Clear_Air_Turbulence

Strauch, B. (2004). *Investigating Human Error: Incidents, Accidents, and Complex Systems*. Farnham: Ashgate.

Wong, B.L.W., & Blandford, A.E. (2002). Analysing ambulance dispatcher decision making: Trialing Emergent Themes Analysis. Presented at the Human Factors 2002, the Joint Conference of the Computer Human Interaction Special Interest Group and The Ergonomics Society of Australia, HF2002, Melbourne.

Woods, D.D., Dekker, S., Cook, R.I., Johannesen, L., & Sarter, N. (2010). *Behind Human Error* (2 edition). Farnham: Ashgate.

Innovative multi-sensor device deployment for fighter pilots activity study in a highly realistic Rafale simulator

Julie Lassalle, Philippe Rauffet, Baptiste Leroy, Laurent Guillet, Christine Chauvin & Gilles Coppin
Lab-STICC UMR CNRS 6285, University of South Brittany, Telecom Bretagne France

Abstract

Cardiac and respiration activities are relatively easy to measure and widely used to monitor pilot workload during simulated or real flight. Few studies include electrodermal and pupil diameter measurements probably due to strong operational constraints. These measures are well-known for being sensitive to mental workload. In a flight framework, the addition of electrodermal activity sensors does not complicate the experimental protocol (wristband wearing) whereas pupillary diameter recording requires a much more sizeable device (eye tracker utilization). In the experiment presented in this paper, heart rate, respiratory rate, skin conductance and pupil diameter were collected during simulated tactical flights. The main novelty of the proposed experimental design relates to eye tracking device integration into a highly realistic flight simulation. To cover the entire pilot visual field and prevent measurement loss, a double-tracking design was tested (i.e. combination of two optical pairs). Preliminary analysis overall confirmed the reliability of this experimental setup showing a high quality of measurement. Nevertheless, extra care should be taken for the skin conductance signal that seems particularly sensitive to movement artefacts. Owing to the observed reliability of data acquisition from the eye tracker it may be possible to extend the proposed device to ocular behaviour measures (scanpaths) in highly realistic flight simulation.

Introduction

The current evolution of aeronautical systems towards unmanned solutions (UAVS, UCAV) brings the place of the human operator in these systems to the foreground. The TAPAS project (stands for Technique d'Analyse pour le Partage d'Autorité dans les Systèmes des systèmes /Analysis Techniques for Shared Authority in the Systems of systems) is a French project between Dassault Aviation, Telecom Bretagne and University of South Brittany. It aims at developing a method for analysing and evaluating different configurations of Human-Human collaboration to enhance the reliability of Human-System relationship. One of the main challenges of this approach is to understand the potential limitations of using these highly autonomous future systems and to define new design principles. The originality and ambition of TAPAS mainly lies in the development of an innovative method, strongly focused on human factors (workload) and related to a design process of new drone control systems.

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Proposed method: operationalization in two stages

Two main steps have been required to develop the TAPAS method. The first one (Guerin et al., 2014) consists in the pilot task analysis (for Navy Rafale aircraft) through intrapatrol radio communications (controller included). These communications were extracted from an air-to-air mission run by an experienced pilot (4 ship lead) at the simulation centre. The task analysis (allo-confrontation method, Mollo & Falzon, 2004) has been made with the help of a Subject Matter Expert (Lt.-Col., French Air Force attached to Dassault Aviation). As a result, twenty nine communication sequences have been identified (such as take-off, tactic flow, fence-in, etc.) to describe collaborative tasks of the two ship lead. This was an essential first step to analyse pilots' activities during flight.
Pilot activity is often studied in terms of mental workload induced by the different flight phases and measured through physiological indicators of autonomous nervous system activity. The second step consisted in the deployment of an experimental setup devoted to on-line recordings (i.e. continuous measurements during the whole flight session) of pilot physiological activity within a highly realistic simulation environment. It should be noted that experiments have taken place during actual training sessions on a Rafale simulator operated by the French Navy. The experimental design had to meet a number of major constraints: (a) to adapt to the simulation environment, (b) not to disrupt pilot activity (unnoticed devices), (c) to allow obtaining high quality data (coverage, reliability).
The main objective of the second step –and to a great extent of the whole developed method – was to detect the critical communication sequences (i.e. increasing mental workload) according to their effect on the physiological activity pattern of the pilots during flight. These sequences can potentially have a negative impact on the success of the flight session. By following critical sequence detection, it will be possible to recommend adaptation of the current communication model between operators and highly autonomous systems.

Pilot activity: contribution of psychophysiological measurements

A lot of studies show the relevance of physiological measurements to monitor pilot activity. A higher physiological activation (activation of the sympathetic branch of the autonomous nervous system) is observed between the resting and flight phases (Karavidas et al., 2010; Lehrer et al., 2010; Veltman & Gaillard, 1996a; 1996b; 1998; Veltman, 2002; Wilson, 2002a; 2002b; Yao et al., 2008; Ylonen et al., 1997) and during the most difficult flight segments namely take-off or approach segments (landing, touch and go) with a high information load. Increased heart rate (Hankins & Wilson, 1998; Lee & Liu, 2003; Veltman & Gaillard, 1996a, 1996b; Yao et al., 2008; Ylonene et al., 1997; Wilson, 2002a; 2000b), respiratory rate (Karavidas et al., 2010; Yao et al., 2008), skin conductance (Wilson, 2002a), pupil diameter (Dehais et al., 2008) and a decrease in the heart rate variability (Hankins & Wilson, 1998; Veltman & Gaillard, 1996b; Wilson, 2002a; Wilson et al., 1994) are reported. The measurement of respiratory rate (RR), heart rate (HR) and heart rate variability (HRV) to study changes in the pilot's mental workload is very commonly used (Casali & Wierwille, 1984; Hankins & Wilson, 1998; Karavidas et al., 2010; Lehrer et al., 2010; Veltman, 2002; Veltman & Gaillard, 1996a, 1996b; Wilson,

2002a; Wilson et al., 1994). On the contrary, very few studies in the aviation field, in simulated or actual flight, report skin conductance (SC) or pupil diameter measurements although they are widely used to study individual mental workload. This lack could be explained by strong operational constraints. SC is conventionally measured via electrodes located at fingertips (a high density site of eccrine sweat glands causing variations in electrodermal activity). However, this configuration cannot be applied for flight context where the presence of electrodes on the fingertips would be inconvenient for pilot activity. Several studies have recently shown that a wrist location (distal inner surface) is an acceptable alternative (Poh et al., 2010; van Dooren et al., 2012). This location expands SC measurement to a broader range of situations including those for which the presence of fingers sensors constitutes an obstacle to the performed activity.
Pupillary changes provide additional information on pilot physiological activity. This measurement is commonly known to reflect the information processing load (Kahneman, 1973; Klingner et al., 2008; Porter et al., 2007). Nevertheless, measurement of eye activity during flight is today primarily studied through the frequency and duration of eyelid blinking. But these indicators reflect the visual load variations more than the mental workload. A consensus exists (Hankins & Wilson, 1998; Stern et al., 1994; Veltman & Gaillard, 1996a) to say that eye blink measurements are specifically sensitive to the amount of visual information to be processed (visual load). Electrooculography (EOG) technique (typically: applying a pair of electrodes around the subject's eyes) is generally used to gather ocular activity (Hankins & Wilson, 1998; Lehrer et al., 2010; Veltman, 2002; Veltman & Gaillard, 1996a; 1998; Wilson, 2002a; Wilson et al., 1994). However, EOG has some limits such as intrusiveness or discomfort (constraints on head movements, trouble with wearing a helmet, etc.) and restricts information collected as part of analysis of pilot activity. For example, pupillary diameter or visual scanpaths cannot be measured. Collecting this information yet appears highly relevant to obtaining the most accurate picture of pilot activity during flight. Integration of a device for measuring pupil diameter and more generally ocular activity in a highly realistic flight simulation is currently a real challenge.

This paper focuses on experimental design operationalization. Added-value of proposed experimental design mainly concerns an eye tracking device used to gather pupil diameter.

Material and Method: deployment of an innovative experimental device

The designed setup makes possible the measurement of the pilot's activity by the means of physiological and ocular indicators in a highly realistic simulation environment. The whole protocol should respect usual training flight conditions without causing troubles for the pilot while allowing optimal measurements.

Subjects

Experiments were conducted during tactical flight training of five male pilots, ages 29-32, to achieve a section lead test. All of them were French Navy fighter pilots. The total piloting experience of participants ranged from 700 and 1100 h with an

average of 870 h and between 150 and 500 h with a mean of 338 h regarding Rafale flight hours.

Simulator

Experiments were performed on a tactical Rafale simulator (see fig. 1) located at the Rafale Simulation Centre at Landivisiau Navy Air Base, France. The cockpit simulator was identical both in appearance and functions to a real Rafale aircraft (real flight instruments and G-seat). Eight retro-projected facets (Apogee 6 Sogitec) arranged in a pseudo-sphere provide a high visual definition. During simulation, the cockpit was placed in the pseudo-sphere allowing a large field vision (330 ° horizontal, 130 ° vertical). The pilot can communicate during the session with his wingman (installed in the same simulator, in a side room) and controller (instructor's room).

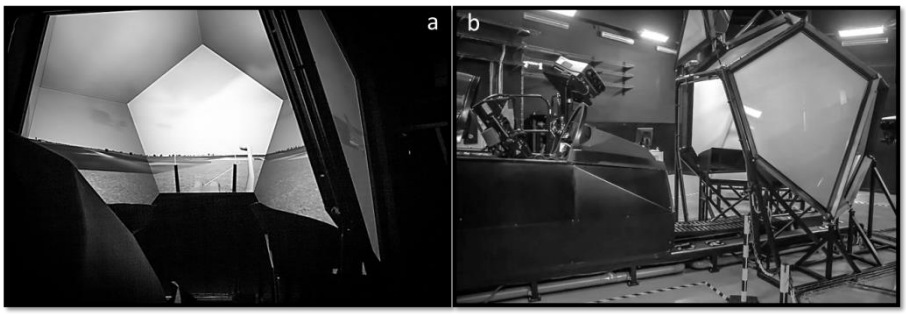


Figure 11. Pseudo-sphere (a) and Rafale simulator presentation (b).

Apparatus

In situ pilot activity was studied using a set of physiological indicators continuously recorded throughout the training session. Heart rate (HR), breathing rate (BR), skin conductance (SC) and pupil diameter (PD) were collected. The sampling frequency was 32 Hz for SC, 250 Hz for HR, 25 Hz for BR and 60 Hz for PD. The cardiac and respiratory activities were measured from a BioHarness3™ belt worn directly on the skin (adjustable elastic strap) around the rib cage just below the chest. The belt integrates a set of sensors for measuring heart rate (electrocardiogram) and respiratory (pressure sensors that detect the expansion of the chest related to respiratory activity). The belt also includes sensors for measurement of acceleration (movements and posture). To fit with experimental field constraints, the SC measurement was achieved by using the Q-Sensor tool (V2) from Affectiva™. The measurement was performed by applying two Ag/AgCl electrodes on the wrist (internal distal face) held by a strap (wristband). The tool also records skin temperature (data control) and acceleration. The latter data can characterize to some extent the physical activity of individuals. Cardiac, respiratory and skin conductance data were locally recorded (i.e. no wireless transmission but device storage, $\geq$ 24h). All sensors (belt and wristband) were installed on pilots prior to the simulated training mission.

The main innovation of the proposed experimental protocol is based on the device to measure pupil diameter, reducing the number of sensors affixed to the same subject and device intrusiveness. One of the main difficulties was to obtain and guarantee a maximal coverage area over the flight to ensure tracking maintenance despite the pilot's head movements. For this, a Double-Tracking Device (DTD) was elaborated. The DTD consisted in the association of two faceLab^{TM} eye trackers (two optical pairs) mounted on a specific support to be easily attached or removed, directly behind the head-up display inside the cockpit (see fig.2). The device (support and DTD) was thought to integrate a simulation environment without causing any inconvenience for the pilot. Furthermore this configuration is supported by the software FaceLab^{TM} Link which generates a virtual tracking device from the two physical eye trackers by merging their data.

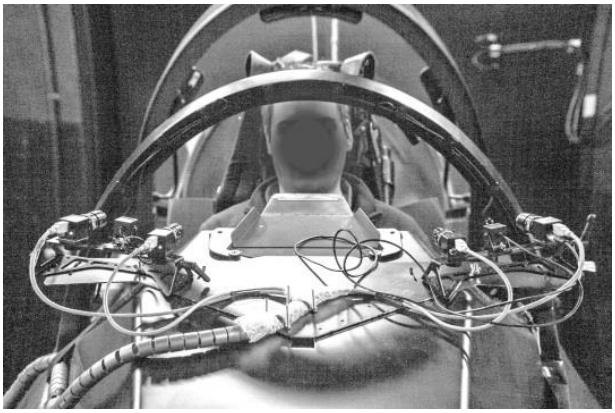


Figure 12. Double-Tracking Device site.

Audio recording (microphone fixed on the pilot's flight suit) and video recording (webcam attached to each side of the cockpit seat) were also collated throughout the training session. These data were required for the subsequent synchronization of physiological and eye data with the flight session timeline. Synchronization is obtained by deleting all the sensor data before the start time of a training mission. Figure 3 shows the complete experimental setup.

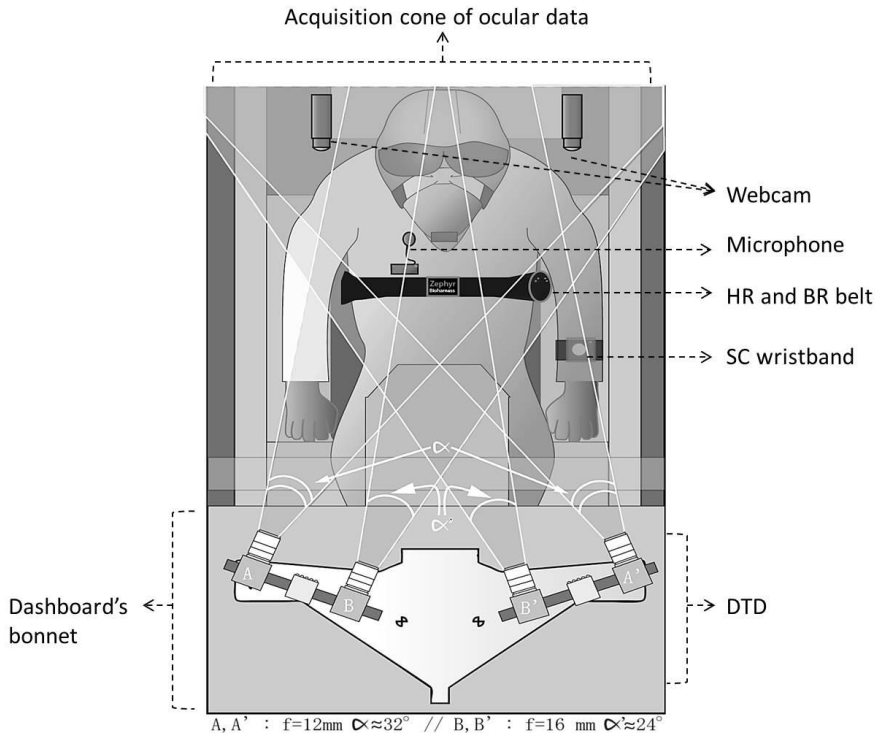


Figure 13. Experimental setup for measuring pilot activity during flight including audio (microphone) and video (webcam) recordings, HR and BR measurements (belt), SC measurements (wristband) and the double-tracking device (DTD) to measure pupil diameter.

Study of device validity

Measurements from two pilots had to be rejected due to technical problems (difficulty of data synchronization) or signal quality. Thus, analyses were performed using data collected from three pilots. The following analyses were conducted using data from six primary simulated tactical flight sessions realized by the three validated pilots. The flight session time period alone has been considered to constitute the analysis data set (data gathered during installation, calibration and sensor removal phases were excluded from the analysis data set).

The proposed device has to be the least intrusive and uncomfortable as possible for pilots. To this end, "contactless" technologies (eye tracker Facelab^{TM}) and unusual sensors (PD, SC) or their unusual location (wrist location) for the study of pilot activity have been preferred and deployed. This kind of device has never been tested. Thus, the first objective was to verify the setup quality according to its data acquisition – i.e. physiological coherence and relevance, data loss quantity (e.g. head movements) or the presence of artefacts. The quality of data acquisition was studied for all the collected measurements:

- Ocular activity (analysis of the pupil diameter data, in mm). It should be noted that eye blinks have been considered for the detection of marginal pupil diameter values,
- Cardiac activity (analysis of R-R intervals computed by the BioHarness3^{TM} software from the electrocardiogram signal, in mV),
- Respiratory activity (analysis of B-B intervals computed by the BioHarness3^{TM} software from the respiratory signal, in mV),
- Electrodermal activity (analysis of the skin conductance, in µS).

Signal quality indicators

Only ocular activity measurements have a quality of acquisition indicator provided by the eye-tracker supplier. This gaze quality indicator ranges from 0 (null quality i.e. no data logged) to 3 (optimal quality of the measurement). To overcome the lack of quality information for the other signals (HR, BR and SC), new indicators were calculated.

First, two signal filters were computed with *Matlab* ® software (see fig. 4):

- *Outliers identification filter*: to count the marginal physiological values from the raw sample,
- *Steps identification filter*: to count the marginal variation between two consecutive data. For cardiac activity and SC, indicators were adapted from Storm et al. (2000) – maximal relative difference of 25% between two R-R intervals - and Sami et al. (2004) – minimal SC value at 2 µS, and maximal temporal slope limited to 2 µS/s -. It should be noted that the skin conductance signal value measured at the wrist is weaker than the classical finger value: a minimal value has thus been visually estimated for each signal.

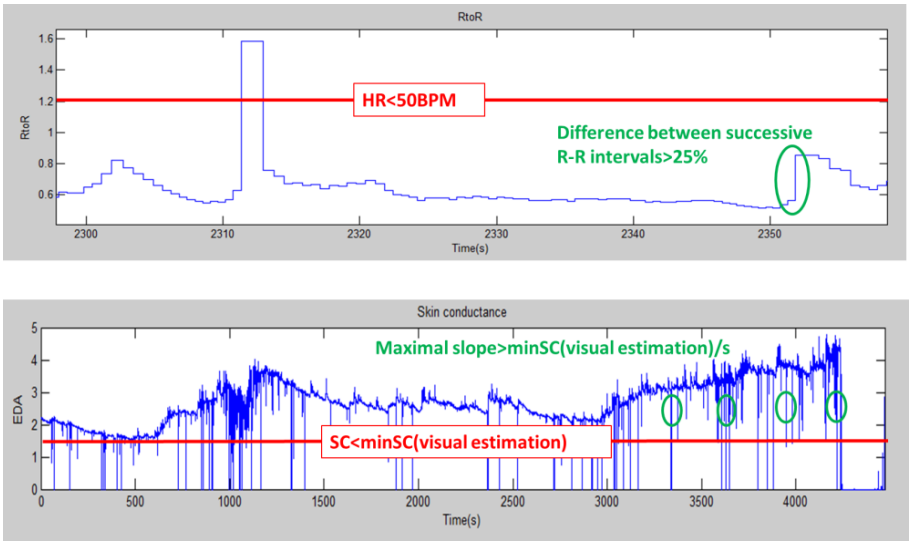


Figure 14. Outliers and steps identification.

Signal quality indicators have been computed to report data acquisition quality on an entire flight session: to enable a comparison, these different indicators were all normalized on a scale ranging from 0 (no valid data) to 100 (all valid data). All these indicators are detailed in the following table.

Table 13. Raw signal processing and indicators computation for Pupil Diameter (PD), Skin Conductance (SC), cardiac (RR intervals) and respiration (BB intervals) activities

Raw data	Outliers and steps identification	Signal quality indicators
Facelab *Pupillary diameter PD* (in mm)	**Outliers** $2<PD<8mm$	*MeanQualiPD=Mean(QualiDiam)* with QualiDiam : *gaze quality indicator ranging from 0 (no data logged) to 3 (optimal quality)*
		$RatioPD = 100 * (1 - \frac{Nb\ Cleaned\ PD - Nb\ Eyeblinks}{Total\ PD\ SaS})$ with cleaned PD: *PD measurements without outliers nor null Qualidiam*
Affectiva *Skin conductance SC* (en µS)	**Outliers** $SC > minSC$ **Steps** $Max\_slope = \lvert SC(k+1) - SC(k) \rvert > \frac{minSC}{sampling\ rate}$ with sampling rate : *number of acquired data by second*	$RatioSC = 100 * (1 - \frac{Nb\ Cleaned\ SC}{Total\ SC\ SaS})$ with cleaned SC : *SC measurements without outliers nor steps*
BioHarness3 *RR and BB intervals* (in s)	**Outliers** $50 \le HR = \frac{1}{RR} \le 240\ BPM$ **Steps** $Step\_RR = \frac{\lvert RR(k+1) - RR(k) \rvert}{RR(k)} > 25\%$	$RatioHR = 100 * (1 - \frac{Nb\ Cleaned\ HR}{Total\ SC\ SaS})$ with cleaned SC : *HR measurements without outliers nor steps*
	Outliers $7 < BR = \frac{1}{BB} < 60\ BPM$	$RatioBR = 100 * (1 - \frac{Nb\ Cleaned\ BR}{Total\ SC\ SaS})$ with cleaned BR : BR *measurements without outliers*

In this table, BPM means Beats Per Minute for heart rate (HR) or Breaths Per Minute for breath rate (BR). "SaS" is used for "Sample Size" and "Nb" for "Number of". "Qualidiam" refers to the FaceLab quality indicator whereas "QualiPD" qualifies the *Matlab*® one.

Device reliability

The following table 2 presents distribution features of the different quality indicators (N=6 flight sessions). An analysis of the homogeneity of the indicators on all the flight sessions has also been conducted by computing the Relative Standard Deviation (RSD=standard deviation/mean). Homogeneity and thus repeatability of the data acquisition (over all the missions) can be questioned if RSD exceeds 15%.

Table 2. Distribution features of quality indicators

Indicators/100	Mean	Standard Deviation	Minimum	Maximum	RSD (%)
RatioQualiDiam	75.34	7.81	65.28	85.26	10.37
RatioPD	71.70	8.16	63.27	82.96	11.39
RatioHR	99.48	0.24	99.17	99.83	0.24
RatioBR	89.74	5.37	82.80	96.46	5.98
RatioSC	56.25	42.03	3.02	94.40	74.72

This table emphasizes three main observations. First, signal quality from HR/BR sensor is excellent (more than 99% of HR signal and 89% of BR signal are physiologically valid) and repeatable (RSD <6%). Second, the quality of PD measurements reaches 72% (RatioPD) of physiologically valid data despite a very constrained activity context, broad head movements and an open visual field. Moreover, the physiological validity filter of pupil diameter computed with *MatLab* (filter defined for the current experiment) and the proposed gaze quality filter proposed by FaceLab (named here as QualiDiam) overlap entirely (i.e. when 2<Ø<8, thus QualiDiam = 3). Additionally, QualiDiam means and the ratio of noisy measurements from the eye tracker are highly correlated (r^2=0.96) and therefore can be used indifferently. Third, quality of SC signal acquisition is lower than quality obtained for the other signals. Indeed, a repeatability problem from one to another flight session (RSD>74%, and RatioSC varies from 3 to 94%) was observed. A visual study of SC signal has shown very noisy graphs for 2 sessions (with the same pilot) with a low RatioSC (<8%).

Discussion

Methodological contributions and perspectives: an innovative experimental setup reliable for high realistic simulation

This paper details an innovative experimental setup to monitor ocular and physiological activity of fighter pilots in a highly realistic environment. The validity and reliability of the setup have been analysed through the quality aspects of data acquisition.

Precisely, the setup enables a high acquisition quality (low level of outliers and steps) and repeatability (RSD<15%) of cardiac and respiratory data (Bioharness™). However, skin conductance measurements (Affectiva™) are to be considered with caution owing to a very noisy signal probably explained by movement artefacts. Despite a valid wrist sensor location, movements of arms and elbows due to pilots' manoeuvres could affect signal quality. The issue of sensor location laterality then arises. In this study, the SC sensor was predominantly affixed to the left wrist and it is interesting to note that the left hand is the most active during Rafale flight. A future study could impose a systematically right location in order to study possible limitations of movement artefacts on the SC signal. Moreover, SC measurements recorded for one of the three pilots systematically presented a poor quality. Excessive perspiration could explain this phenomenon by either generating numerous outlier data or leading to artefacts due to sensor movements (sweat can

lead to wristband slips). An ankle location seems to be a better alternative. Such a device could be studied in future research.
Furthermore, reliable data have been collected with DTD with a large cone of acquisition (full cockpit). The DTD reliability observed for pupil diameter suggests that this utilization could be extended to ocular scanpath collection for studying fighter pilot activity. This scope brings an interesting field of research perspectives which is not much investigated nowadays for an *in situ* flight environment framework. Therefore, a further step could consist in DTD optimization by testing its capacity to gather usable ocular behavioural measurements with the same level of quality.
Overall, results on the DTD and the whole setup indicated an effective and responsive device. It was successfully deployed and offers an ambulatory and non-invasive solution for a realistic flight environment to gather high quality data without affecting pilot activity. In the long term, the DTD and the detailed whole setup could be adapted for real flight deployment.

Practical contributions and perspectives: integration in TAPAS method for assessing mental workload in the context of Human-System collaboration

In addition to the assessment of the validity of the apparatus, the propositions presented in the paper contribute to the development of the TAPAS two-stage method (based on task and activity analyses).
Thus, the *Matlab* routines for identifying signal outliers and steps could also be used to clean the data vectors for preliminary signal processing. Indeed, this is necessary to calculate and then to compare the accurate mean values of physiological data on different flight sequences, and therefore to classify communication and activity sequences according to a level of mental workload.
To conclude, this contribution supports the processes of physiological activity data processing (dotted lines) in Figure 4. This figure illustrates the global TAPAS method for assessing mental workload in the context of Human-Human and Human-System collaboration.

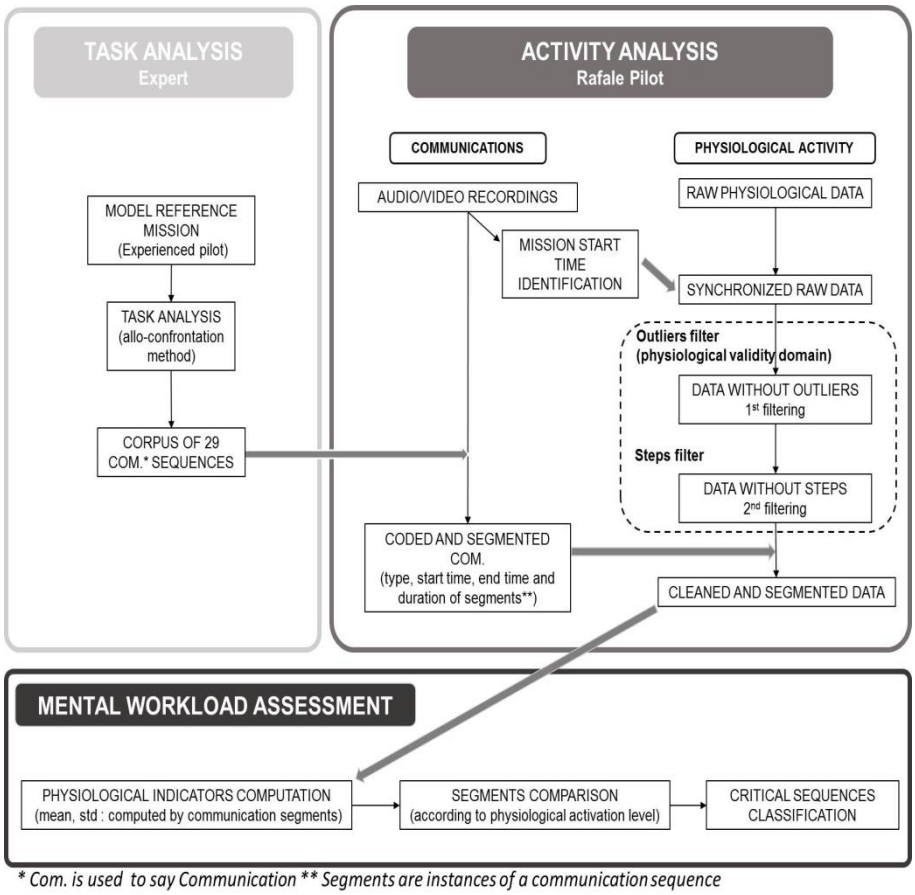


Figure 15. Global TAPAS method.

Acknowledgements

TAPAS project would not have been possible without the approval of the Navy air base at Landivisiau, France) and the assistance of the simulation centre team. The authors also wish to thank all fighter pilots who agreed to participate in this project.

References

Casali, J.G., & Wierwille, W.W. (1984). On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics, 27*, 1033-1050.

Dehais, F., Causse, M., & Pastor, J. (2008). Embedded eye tracker in a real aircraft: new perspectives on pilot/aircraft interaction monitoring. In *3rd International Conference on Research in Air Transportation Proceedings*. Fairfax, USA: Federal Aviation Administration.

Guerin, C., Leroy, B., Chauvin, C. & Coppin, G. (2014). Task analysis from the expert point of view: a prerequisite condition to analyse physiological

activity of fighter pilot aircraft. Poster presented at *HFES Europe Chapter 2014 Annual Conference*, Lisbon, Portugal.

Hankins, T.C., & Wilson, G.F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation, Space, and Environmental Medicine, 69*, 360-367.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Karavidas, M.K., Lehrer, P.M., Lu, S.E., Vaschillo, E., Vaschillo, B., & Cheng, A. (2010). The effects of workload on respiratory variables in simulated flight: A preliminary study. *Biological Psychology, 84*, 157-160.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *ETRA 2008 Proceedings* (pp. 69-72).New York, USA : ACM.

Lee, Y.H., & Liu, B.S. (2003). Inflight workload assessment: Comparison of subjective and physiological measurements. *Aviation, Space, and Environmental Medicine, 74*, 1078-1084.

Lehrer, P., Karavidas, M., Lu, S.E., Vaschillo, E., Vaschillo, B., & Cheng, A. (2010). Cardiac data increase association between self-report and both expert ratings of task load and task performance in flight simulator tasks: An exploratory study. *International Journal of Psychophysiology, 76*, 80-87.

Mollo, V., & Falzon, P. (2004). Auto- and allo-confrontation as tools for reflective activities. *Applied Ergonomics, 35*, 531–540.

Poh, M.Z., Swenson, N.C., & Picard, R.W. (2010). A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *Biomedical Engineering, 57*, 1243-1252.

Porter, G., Troscianko, T., & Gilchrist, I.D. (2007). Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology, 60*, 211-229.

Sami S., Seppänen M., & Kuusela A. (2004). Artefact correction for heart beat interval data. In *1st Probisi 2004 Proceedings*. Jyväskylä, Finland: University of Jyväskylä.

Stern, J.A., Boyer, D., & Schroeder, D. (1994). Blink rate: a possible measure of fatigue. *Journal of the Human Factors and Ergonomics Society, 36*, 285-297.

Storm H., Fremming A., Odegaard S, Martinsen O, & Morkrid L. (2000). The development of a software program for analyzing spontaneous and externally elicited skin conductance changes in infants and adults. *Clinical Neurophysiology*, *111*, 1889-1898.

Van Dooren, M., de Vries, J., & Janssen, J.H. (2012). Emotional sweating across the body: Comparing 16 different skin conductance measurement locations. *Physiology & behavior, 106*, 298-304.

Veltman, J.A. (2002). A comparative study of psychophysiological reactions during simulator and real flight. *The International Journal of Aviation Psychology, 12*, 33-48.

Veltman, J.A., & Gaillard, A.W.K. (1996a). Physiological indices of workload in a simulated flight task. *Biological Psychology, 42*, 323-342.

Veltman, J.A., & Gaillard, A.W.K. (1996b). Pilot workload evaluated with subjective and physiological measures. In K. Brookhuis, C. Weikert, J.

Moraal, and D. de Waard (Eds.), *Aging and Human Factors, Proceedings of the Europe Chapter of the Ergonomics Society.* (pp.107-128). Haren, The Netherlands : Traffic Research Centre, University of Groningen.

Veltman, J.A., & Gaillard, A.W.K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics, 41*, 656-669.

Wilson, G.F., Fullenkamp, P., & Davis, I. (1994). Evoked potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. *Aviation, Space, and Environmental Medicine, 65*, 100-105.

Wilson, G.F. (2002a). An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology, 12*, 3-18.

Wilson, G.F. (2002b). A comparison of three cardiac ambulatory recorders using flight data. *The International Journal of Aviation Psychology, 12*, 111-119.

Yao, Y.J., Chang, Y.M., Xie, X.P., Cao, X.S., Sun, X.Q., & Wu, Y.H. (2008). Heart rate and respiration responses to real traffic pattern flight. *Applied psychophysiology and biofeedback, 33*, 203-209.

Ylonen, H., Lyytinen, H., Leino, T., Leppaluoto, J., & Kuronen, P. (1997). Heart rate responses to real and simulated BA Hawk MK 51 flight. *Aviation, Space, and Environmental Medicine, 68*, 601-605.

Is simulation (not) enough? Results of a validation study of an autonomous emergency braking system on a test track and in a static driving simulator

Martin Jentsch & Angelika C. Bullinger
TU Chemnitz
Germany

Abstract

Comparison of data gathered with real vehicles and with a driving simulator is still heavily debated. This paper provides results of a validation study with 164 participants who tested an autonomous emergency braking system (AEBS) either in a driving simulator or on a test track. Participants were similar concerning age and driving experience and experienced real driving on a test track and in a 180° Field of View (FOV) static driving simulator. Study design, scenarios and questionnaires to assess e.g. drivers' perceived degree of dangerousness of the situation, perceived usefulness of the system in each scenario and overall acceptance were used in both set ups. Additionally, vehicle dynamics were recorded. Participants drove one of six types (three braking intensities each with two different times for acoustical warnings) of the system. Three traffic scenarios (e.g. distracted driver with a sudden braking of the leading vehicle) with a moving vehicle ahead and two scenarios with a stationary target (e.g. AEBS intervention during evasive manoeuvre) were accomplished by each participant. It was found that participant's reaction in the simulator is comparable to the reaction on the test track. Participants' judgment of the system, situation and overall acceptance could be shown to be almost the same.

Introduction

In the last 20 years Advanced Driver Assistance Systems (ADAS) developed and diffused rapidly. Despite their benefits, it cannot be ignored that the driving task can seriously change if driving with ADAS support. One example is the additional need of the continuous monitoring of the ADAS (Spanner-Ulmer, 2008). For this reason, it is necessary to ensure that the driver quickly understands the function and the boundaries of the system and is able to operate it safely. This is necessary to ensure that the driver does not put himself or other road users at additional risk, for instance by misinterpreting the ADAS. If these prerequisites are fulfilled the driver is more likely to accept the system, is willing to use it continuously and the ADAS can reach its full potential to increase traffic safety or drivers' comfort. When developing ADAS, the manufacturers have to face the challenge to design the systems according to the driver's needs and to ensure technical and functional reliability. To determine suitable specifications of a new ADAS, requirements are usually obtained by studies

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

with participants, ideally the future customers of the company, who experience and evaluate the system in early development stages. These experiments can be performed in real road traffic as a "field" or "on road" test, on a test track or in a driving simulator.

Surprisingly, there is little knowledge to which extent results from experiments with ADAS are comparable between driving simulators and real vehicles. This refers on the one hand to possibly modified (driving) behaviour (objective measures of the vehicle dynamics and the drivers' behaviour) in the driving simulator. On the other hand, the subject's assessment and the overall acceptance (subjective measures) of the ADAS may be different in the driving simulator. Furthermore it is uncertain whether the relations between objective and subjective measures are influenced by the test environment.

This leads to the research questions:

- Is it possible to get similar findings concerning driving/driver's behaviour, system evaluation/overall acceptance and situation evaluation in a static driving simulator compared to findings from a real vehicle for interventions of an autonomous emergency braking system (AEBS)?
- Which objective and subjective measures are suitable for ADAS evaluation in a static driving simulator?

If the feasibility of an experimental procedure in the static driving simulator can be demonstrated this will, of course, not entirely substitute tests with real vehicles. Functional reliability of the ADAS and a final subject assessment will always be necessary in a real vehicle. However, it would be possible to perform experiments in early concept or very early development phases of the ADAS without the need of a fully operative ADAS in a real vehicle. This allows important insights that are valuable for the design of the system, which can positively influence the development process. When optimization potentials regarding the ADAS specification are found as early as possible, development times can be shortened and development costs can be reduced.

In this paper, the validity of a static driving simulator for experiments with ADAS, which actively intervene in the longitudinal control of the car, will be examined. Therefore the ADAS *Aktive Gefahrenbremsung*, an Autonomous Emergency Braking System (AEBS), which was developed in the *AKTIV*[3] research initiative, was selected as example. The AEBS enables autonomous prevention of rear-end collisions without driver's action. Therefore it is representative for similar ADAS, which actively intervene in the driving task and systems, which will enable highly or fully automated driving in the future.

[3] *AKTIV* was funded by the Federal Ministry of Economics and Technology (BMWi). TU Chemnitz did the evaluation of the *Aktive Gefahrenbremsung* on a test track as part of a subcontract.

State of the art

The driving task (Donges, 1982; Bubb, 2003) represents a highly complex task for the driver in which he may reach his limits concerning human perception and reaction. Traffic accidents can be the consequence of a non - or only bad performed - completion of the driving task (Lee, 2008). Accidents in longitudinal traffic, such as rear-end collisions, account for a percentage of about 25 - 30% (Hannawald, 2013) of all traffic accidents. This illustrates the high potential for safety ADAS, especially in longitudinal control, such as AEBS. Particularly for actively intervening ADAS it is crucial to know and ensure the driver's interaction with the vehicle and the ADAS. This is even more important for systems which intervene at higher speeds to prevent misuse, abuse and associated negative effects of the ADAS relating to road safety (Knapp et al., 2009).

Measuring driver and system performance

Characteristic values to evaluate the driver's interaction with ADAS can be divided into physically measurable, objective measures and subjective measures, which are obtained by interviewing the driver, e.g. using questionnaires.

Physically measurable, objective measures can be subdivided into vehicle dynamics measures and driver behaviour measures (Wierwille et al., 1996; Johansson et al., 2004; Östlund et al., 2005; Dotzauer et al., 2011; Dettmann, 2012). From the recorded raw data further values such as minima, averages or maxima can be calculated within specified measurement intervals in order to derive results concerning the desired research question.

Subjective measures can be divided into measures of acceptance (Arndt & Engeln, 2008), system evaluation (Riedel & Arbinger, 1997) and situation evaluation (Kiefer, Flannagan & Jerome, 2006). For ADAS that are not on the market and therefore cannot have been experienced by drivers yet measuring acceptance is difficult. In the experiment drivers experience a new ADAS for the first time and only over a limited period of time. Referring to the theory of planned behaviour (Ajzen, 1991), the attitude toward the behaviour, the intention (to use the ADAS) and the perceived system's characteristics can be good predictors of the future driver's acceptance.

With questionnaires, it is possible to let the driver assess the perceived usefulness and usability (Fastenmeier & Gstalter, 2008) or the overall satisfaction (Pataki, 2005) of an ADAS. Since ADAS that focus on increasing traffic safety can only be experienced in complex or hazardous traffic situations, the situation evaluation is closely linked to the system evaluation. The perceived driving situation can be measured by participants' perceived danger of the situation, the characteristics of the situation, e.g. concerning crucial object in the scenario or the estimation of distances. Measures of acceptance are giving developers insights about the driver's attitude towards the ADAS and provide an estimation of his actual will to use the system in real traffic.

Based on these considerations a set of the most frequent and according to the literature most promising objective measures to assess systems' and driver's performance for AEBS was chosen and questionnaires were designed for the experiments (see table 1).

Table 1. Objective and subjective measures for the validation study

<table>
<tr><th>Objective measures</th><th>Subjective measures</th></tr>
<tr><td>Vehicle dynamics</td><td>Acceptance</td></tr>
<tr><td rowspan="3">▪ Speed
▪ Longitudinal acceleration
▪ Distance (incl. Time Headway (THW) & Time to Collision (TTC)
▪ Brake Reaction Time (BRT)
▪ Pedal Measures / pedal activity
▪ Steering behaviour/ steering wheel angle</td><td>▪ Attitude toward the behaviour
▪ Intention (to use the ADAS)
▪ Perceived system's characteristics</td></tr>
<tr><td>System evaluation</td></tr>
<tr><td>▪ Usefulness
▪ Usability
▪ Overall satisfaction</td></tr>
<tr><td>Driver behaviour</td><td>Situation validation</td></tr>
<tr><td>▪ Glance behaviour</td><td>▪ Danger
▪ Objects in a situation
▪ Distances in a situation</td></tr>
</table>

Issues regarding the driver-vehicle-interaction during the intervention of an AEBS and acceptance towards the system cannot be answered in real road traffic. The main reason for that is that these systems can only be experienced in perilous situations. This causes a far too big threat to the safety of the participant and other road users which disqualifies the test environment "real road" for AEBS experiments in early stages. Therefore, only the test environments test track and driving simulator are suitable.

Test methods

The three main test quality criteria objectivity, reliability and validity are the basic requirements while planning, conducting and interpreting experiments (Bryant, 2000). If the experiment does not take place in the real road traffic, questions concerning the validity of the findings may occur. Two kinds of validity can be distinguished: internal (adequately accurate acquisition of parameters, extent to which a causal conclusion based on a study is warranted) and external (transferability and generalizability of the results). For comparative studies between e.g. a driving simulator and a test track experiment, as conducted in the case at issue for this article, the external validity can be distinguished into two types (Blana, 1996): absolute validity (same or similar measured values between the test environments) relative validity (same effects or rank order but different absolute

values, depending on the test condition between the test environments). By examining the above mentioned measures regarding external validity, assumptions can be made concerning the feasibility of the static driving simulator for AEBS - or similar ADAS - evaluation experiments.

Empirical Study

Function and types of the AEBS

Six different types of the AEBS have been developed by the *AKTIV* research initiative. All of these are designed to avoid accidents within the limits of the system and to brake the vehicle until standstill. The types consist of three braking intensities (full, partial and combined braking) with two different times for acoustical warnings each. The full braking (FB) intervenes until standstill with a deceleration of 7 m/s². In the partial braking (PB), the vehicle decelerates until standstill with 4 m/s².

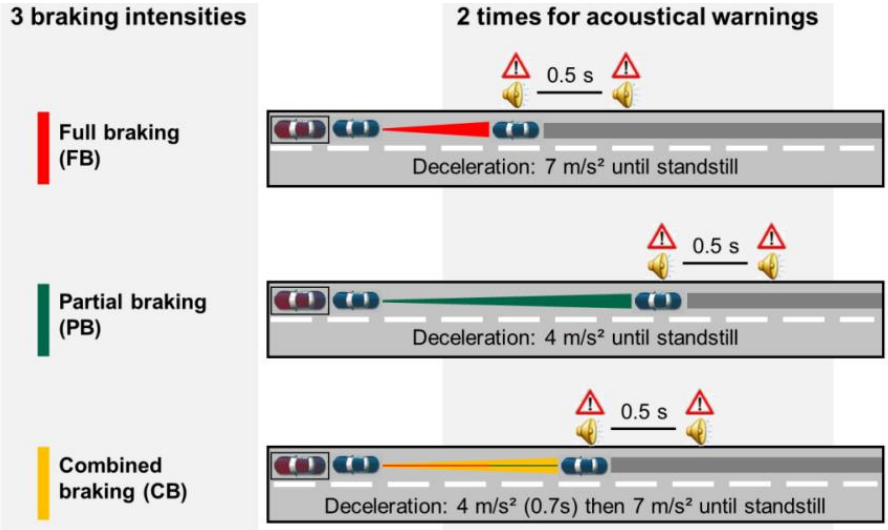


Figure 1. Types of the AEBS grouped by braking intensity and warning times

The combined braking (CB) decelerates with 4 m/s² in the first 0.7 s and afterwards with 7 m/s² until standstill. For each strategy of braking there is an acoustic warning, which starts either simultaneously or 0.5 s before the braking intervention of the AEBS. Apart from the characteristics of each type the factors warning time and braking intensity can be compared. Figure 1 shows the types of the AEBS with their principal modes of operation.

The six types were implemented in two vehicles for the test track experiment. For the presentation of the AEBS in the static driving simulator they were simulated with the software “SILAB 3.0” given identical system characteristics to those in the real vehicles.

The airport and test area Großenhain was chosen as test track. The driving simulator experiment took place in a static driving simulator with 180° projection at the Technische Universität Chemitz (Jentsch, 2014).

Scenarios

Since the AEBS is an ADAS that is currently not available it is of great importance to illustrate the participants the functioning of the AEBS in various scenarios in a fixed order. Not only was a reliable assessment of the acceptance towards the system and the system evaluation by the participants in the focus of the experiment. Also knowledge about the interaction of the driver with the AEBS had to be examined. Therefor it was necessary to define different scenarios with respect to everyday situations where an AEBS would intervene.

Scenario 1 - Unexpected impending frontal collision with visual driver distraction: AEBS intervention to prevent a rear-end collision with a vehicle in front, which suddenly decelerates. A moving target, which decelerates unexpectedly for the participant while he is distracted by a visual-motoric secondary task and driving at 60 km/h, is used to represent the scenario.

Scenario 2 - Stop&Go situation without driver distraction: Scenario 2 represents a classic Stop&Go situation at low speed (maximum 40 km/h). This scenario represents an accident hotspot in longitudinal traffic (Schaller, 2009). It may also illustrate the participant the possibly existing disadvantages of the warning signal before the braking intervention.

Scenario 3 - Announced AEBS intervention without driver distraction: Scenario 3 represents a modification of Scenario 1, where this time the participant is not visually distracted. Additionally a verbal explanation of the AEBS function is given before starting the scenario. The participants should experience consciously the AEBS intervention to get a better understanding of the timing and intensity of braking and acoustical warning.

Scenario 4 - Unexpected AEBS intervention during evasive manoeuvre: Depending on vehicles' velocity and the intensity of deceleration during the braking manoeuvre it is possible that the distance necessary for fulfilling an evasive manoeuvre is shorter than the necessary distance for braking. Problems can occur for AEBS because the system may detect a critical situation and starts intervening. However the driver may plan an evasive or overtaking manoeuvre. Scenario 4 is used to examine this driving situation with a velocity of 65 km/h.

Scenario 5 - Announced AEBS intervention when approaching a stationary obstacle: At the end of the experiment the participants are asked to compare all six types of the system in scenario 5. For this purpose, the participants drive consciously and without visual distraction with a predetermined speed (50 km/h) towards a stationary target. Before the first run in scenario 5, the participants are explained that there are six types of the system without going into their characteristics. A comparative analysis of the subjective assessment of the system between the types of the system can be made with the help of this scenario.

Study Design

Up to scenario 3, a between-subjects-design was chosen and a moving target was used (see figure 2). Each participant consistently drove these scenarios with the same type of the AEBS. By doing so, each participant should get the opportunity to estimate each detail of the system at different speeds and in different situations. The system types were assigned to the participants in a way that every type was driven by the same number of male and female participants and occasional and frequent drivers.

Figure 2. Moving (left) and stationary (right) target on the test track (Jentsch et al. 2012)

For scenario 4 and 5, a stationary target was used (see figure 2). In scenario 4, the between-subjects design of the first three scenarios was generally maintained. Types of AEBS with full braking interventions were excluded from this scenario due to the short necessary distance for braking. Participants with FB types were equally assigned to partial or combined braking intensities in scenario 4. For scenario 5 a within-subjects design was chosen. Participants began with types of the AEBS they already were familiar with from the first three scenarios. This was followed by runs with the other five types. The order was balanced to eliminate position and sequence effects.

Participants

79 people joined the experiment on the test track, 92 took part in the driving simulator. Two participants of the 92 already dropped out during training sessions other five had to end their attendance due to simulator sickness (Reason, & Brand, 1975). In the end data of 85 participants was collected in the driving simulator. Participants were recruited from a database, with flyers and announcements. They were given a reward of 25 € for participating on the test track were the experiment lasted approx. 2,5 h and 15 € in the driving simulator where the duration was approx. 1.5 h.

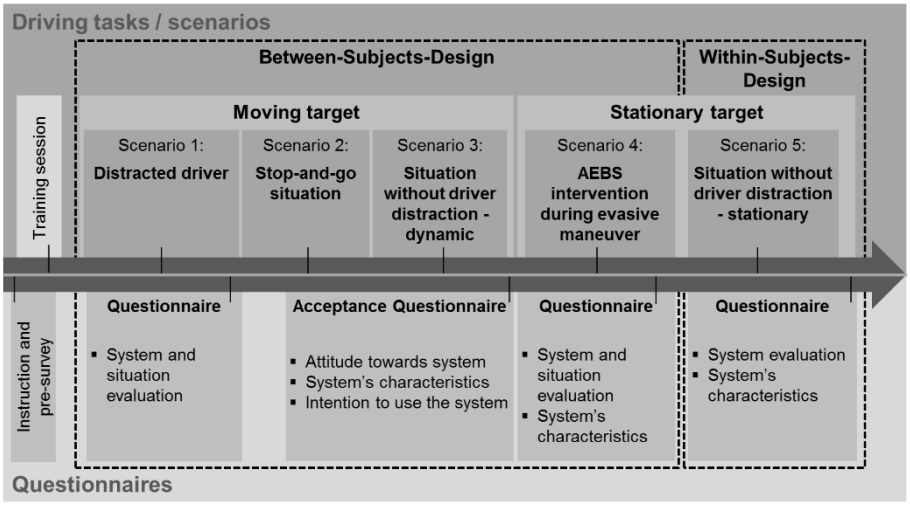


Figure 3. Experimental design

To investigate the influence of driving frequency, groups of occasional (< 10.000 km/year) and frequent drivers (> 15.000 km/year) were formed. Sex of the participants was almost equally distributed within these groups to minimise the influence of the participant on the comparison of the experimental environments the samples were taken such as participants were similar concerning age and driving frequency in both test environments. Table 2 shows participant's characteristics on the test track and in the driving simulator.

Table 2. Age of participants and kilometres driven within the last year

Participant's characterisitics	*Test track*			*Driving simulation*			*t-test*		
	N	*M*	*SD*	*N*	*M*	*SD*	*df*	*t*	*p*
Age [years]	79	28,9	9,51	90	28,7	6,16	167	0,128	,898
km driven last year	79	13678	9903	90	14471	12411	167	-0,455	,650

Data analysis

To answer the questions mentioned in the introduction it is necessary to examine, to what extent a dependency of the (driving) behaviour, the subjective assessment and the relations between the (driving) behaviour and the subjective assessment of the types of the AEBS exists in both test environments. Furthermore the results must be analysed in the context of different driving scenarios and whether there are differences in the (driving) behaviour and the subjective evaluation of the AEBS depending on the annual kilometres covered by the driver. Figure 4 summarises these issues and illustrates the dependent and independent variables for the validation study.

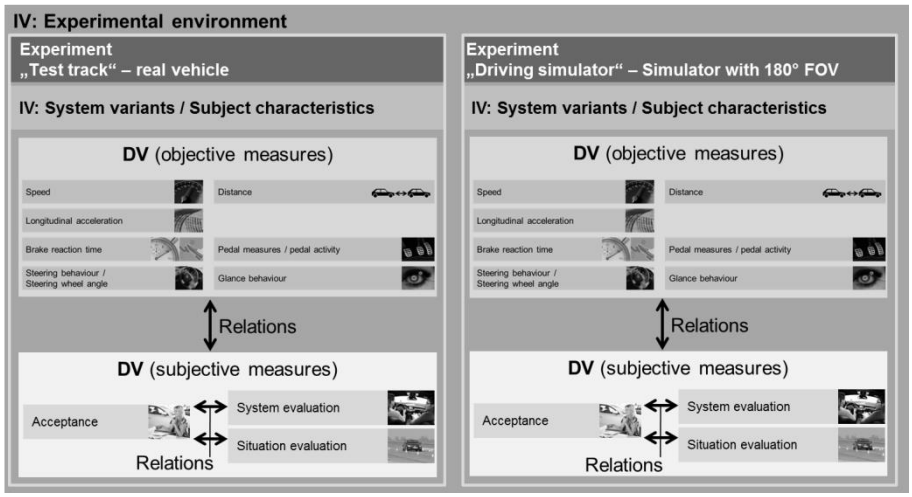


Figure 4. Variable description for the validation study (DV - dependent variable; IV - independent variable)

In order to make a reliable conclusion about the suitability of a static driving simulator for the investigation of an AEBS, it is necessary to measure the dependent variables with an identical experimental setup and a similar sample in both test environments. The influence of the test environment (independent variable) on the measures can then be determined by differences regarding the dependent variables. A comparison of the measures on the test environment makes it possible to determine the absolute validity of the static driving simulator. Within each test environment, braking intensity and timing of the acoustical warning as well as the driver's annual kilometres, divided into two groups (occasional and frequent drivers) are the independent variables. The relative validity is examined by the influence of the independent variables on the measures within one test environment and compared to the other.

Results

Example for analysis procedure

First of all, hypotheses were literature-based formulated and then tested using t-tests, ANOVAs or correlations for all relevant scenarios regarding the measure under investigation. In total 52 hypotheses were formulated for objective and 46 for subjective measures. Furthermore 27 hypotheses focused on the relation between the measures. The analysis procedure will be explained using two of the five hypotheses concerning the objective measure *speed:*

> *H1: Participants will choose lower speed when driving a full braking type of the AEBS, compared to other braking intensity types, in scenario 1. (explained by driver's compensation of shorter following distance when distracted)*

H2: The measure speed does not show differences comparing the two experimental environments for scenario 1, 3 and the 1st run of scenario 5.

The recorded values for speed at the beginning of the manoeuvre for the three braking intensities are shown in table 3.

Table 3. Speed at the beginning of the manoeuvre for all braking intensities

Scenario		Speed [km/h] at beginning of maneuver								
		Full braking			Partial braking			Combined braking		
		N	M	SD	N	M	SD	N	M	SD
Scenario 1	Test track	25	59,26	2,30	24	58,14	1,76	26	59,06	2,89
	Driving sim.	28	62,19	1,36	26	62,15	1,69	31	61,74	2,07
Scenario 3	Test track	22	58,35	1,65	23	58,42	1,96	24	58,05	2,03
	Driving sim.	27	62,82	1,09	26	62,12	1,38	30	62,83	1,23
Scenario 5 (1st run)	Test track	23	46,03	2,51	25	46,14	2,65	25	45,36	3,11
	Driving sim.	25	51,39	2,44	25	50,52	2,31	30	49,98	1,91

Neither on the test track (ANOVA, $F(2,72) = 1{,}555$; $p = ,218$) nor in the driving simulator (ANOVA, $F(2,80) = 2{,}883$; $p = ,062$) data regarding speed at the beginning of the manoeuvre show differences depending on the braking intensity in scenario 1. *H1* is rejected in both experimental environments. Participants are generally driving 3 to 5 km/h faster in the driving simulator than on the test track. *T*-test are proving significant speed differences between simulator and test track for all braking intensities (Full Braking: $t(51) = -5{,}699$; $p < ,001$; Partial Braking: $t(48) = -8{,}220$; $p < ,001$; Combined braking: $t(55) = -4{,}061$; $p < ,001$). *H2* is also rejected.

While relative validity can be confirmed, absolute validity is not given at first sight due to higher values in the driving simulator. Taking into account that speedometers in real vehicles are always indicating a velocity that is higher than the one that is actually driven, this result is not very surprising. For the objective measure speed it can be concluded that speedometers in the driving simulator must use an offset, similar to real vehicles, to gain data in the driving simulator that is showing absolute validity. Taking this into consideration when designing experiments in driving simulators speed can be seen as a suitable measure for evaluating ADAS.

Interpretation of results and conclusion

As a result of the study, insights were gained on the suitability of objective and subjective measures for evaluating ADAS intervening into the longitudinal control of the vehicle in a static driving simulator. The analysis for all measures was similar to the described procedure above. Measures can be distinguished between the ones which are suitable, partly suitable and not suitable. Suitable measures are showing mostly relative and absolute validity or the differences found between the experimental environments can be minimized simply, as shown on the example speed above. Partly suitable measures are showing in most scenarios relative or absolute validity while not suitable measure are mostly not showing relative nor

absolute validity. In table 4 suitable measures are marked in green, partly suitable in yellow and not suitable in red.

Table 4. Summary of suitability of the measures

Objective measures	Subjective measures
Vehicle dynamics	
Speed	**Acceptance**
Longitudinal acceleration	Attitude toward the behaviour
Distance	Intention (to use the ADAS)
THW and TTC at driver's intervention	Perceived system's characteristics
Minimum and average THW	**System evaluation**
Brake reaction time	Moment and usefulness of intervention / acoustical warning and best variant
Pedal Measures	Brake intensity and overall satisfaction
Movement time and number of driver's intervention in perilous situations	**Situation validation**
Maximum of brake pedal actuation und time to reach maximum	Danger, probability of accident caused by ADAS, vehicle control
Pedal activity in non-perilous situations	Probability of accident without the ADAS and evading distance
Steering behaviour	
Steering wheel angle	
Evading reaction	
Driver behaviour	
Glance behaviour	

suitable partly suitable not suitable

The results show that an initial assessment of intervening ADAS in a static driving is possible since subjective measures are mostly suitable or partly suitable. The participants react to imminent collisions in longitudinal traffic in a driving simulator similar to real vehicles. However, especially on the objective measures there are non-negligible differences. Measured *longitudinal acceleration* in the driving simulator within the first 0,5 s after braking intervention is generally lower than on the test track. This can be explained by higher *brake reaction times* in the driving simulator. Participants are also showing higher *minimum and average THW* when following a leading vehicle in the driving simulator (scenario 2). In the driving simulator experiment participants are not showing *evading reactions* during the braking manoeuvre which was frequently observed on the test track. To avoid misinterpretation, these restrictions should be strictly taken into consideration.

A subjective evaluation by the participants allows in the driving simulator a very good assessment of the system's characteristics. The relations between the objective measures and the system and situation evaluation are identical to those on the test track. This implies that not only the results of the questionnaires are similar between the two experimental environments. Also their occurrence, in relation to the actual

behaviour of the participants, can be explained profoundly. There is a higher relation between the overall acceptance and the system and situation evaluation on the test track compared to the driving simulator. Even if acceptance in the driving simulator is not significantly different from those measured on the test track, this can be seen as an indication that the understanding of the system is more consistent on the test track.

Discussion and outlook

An experiment has been designed and conducted on a test track and in a static driving simulator to determine the (driving) behaviour during interventions of system characteristics of an AEBS. The system was experienced and evaluated by 80 participants in each experimental environment. The focus was laid upon chosen objective and subjective measures in order to derive conclusions about their suitability for experiments to evaluate longitudinal intervening ADAS in a static driving simulator.

The experiments have been carried out with unbiased participants. This means that they were unfamiliar with the AEBS. A self-braking vehicle calls out an enormous enthusiasm at first glance. This could be the reason why different system characteristics and possible disadvantages were not recognized by the participants and the acceptance and system evaluation was very positive in both environments. The described study showed that despite the missing haptic feedback in the static driving simulator participants are able to give an evaluation of the system similar to when they are experiencing the AEBS in a real vehicle. This allows incorporating static driving simulators in an early stage of the development process.

Literature

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*, 179-211.

Arndt, S. & Engeln, A. (2008). Prädikatoren der Akzeptanz von Fahrassistenzsystemen. In S. Schade, and A. Engeln, A. (Eds.), *Fortschritte der Verkehrspsychologie* (pp. 313-337). Wiesbaden: GWV Fachverlage.

Blana, E. (1996). *Driving Simulator Validation Studies: A Literature Review*. Working Paper 480, University of Leeds. Leeds, UK: Institute of Transport Studies.

Bubb, H. (2003). Fahrerassistenz – primär ein Beitrag zum Komfort oder für die Sicherheit? In *VDI-Bericht Nr. 1768 (*pp. 25-44). Düsseldorf: VDI-Verlag.

Bryant, F.B. (2000). Assessing the Validity of Measurement. In L.G. Grimm and P.R. Yarnold (Eds.). *Reading and understanding MORE multivariate statistics* (pp. 99-146). Washington, D.C., USA: American Psychological Association.

Dettmann, A. (2012). *Vergleichende Untersuchung von Simulator- und Realversuchen am Beispiel von Blickverhalten und Aktiver Gefahrenbremsung*. Diplomarbeit, TU Chemnitz. Chemnitz: Professur Arbeitswissenschaft.

Donges, E. (1982). Aspekte der Aktiven Sicherheit bei der Führung von Personenkraftwagen. *Automobil-Industrie, 1982-2*, 183-190.

Dotzauer, M., Piccinini, G., Haupt, J., Beggiato, M., Berthon-Donk, V., Wege, C., Bueno, M. Hajek, W., Bhatti, G., & Gouy, M. (2011). *Design of longitudinal studies in driving simulators and real traffic conditions - empirical work to investigate drivers' adaptation processes*. Deliverable 6: ADAPTATION Project. Paris: IFSTTAR.

Fastenmeier, W. & Gstalter, H. (2008). Beitrag psychologischer Modelle und Methoden zur Bewertung von Fahrerassistenzsystemen. *Zeitschrift für Arbeitswissenschaft, 62*, 15-24.

Hannawald, L. (2013). Das Unfallgeschehen in Deutschland und Situationen unsicheren Fahrens. Paper presented at the *6. Darmstädter Kolloquium Mensch + Fahrzeug - Maßstäbe sicheren Fahrens*, 2013, March. Darmstadt.

Jentsch, M. (2014) *Eignung von objektiven und subjektiven Daten im Fahrsimulator am Beispiel der Aktiven Gefahrenbremsung - eine vergleichende Untersuchung* PhD thesis, TU Chemnitz. Chemnitz: Universitätsverlag Chemnitz.

Jentsch, M., Lindner, P., Spanner-Ulmer, B., Wanielik, G., & Krems, J.F. (2012). Nutzerakzeptanz von Aktiven Gefahrenbremsungen bei statischen Zielen. In: Gesellschaft für Arbeitswissenschaft (Eds.), *Gestaltung nachhaltiger Arbeitssysteme, Bericht zum 58. Kongress der Gesellschaft für Arbeitswissenschaft* (pp. 477-481). Dortmund: GfA-Press.

Johansson, E., Engström, J. Cherri, C., Nodari, E., Toffetti, A., Schindhelm, R., & Gelau, C. (2004). *Review of existing techniques and metrics for IVIS and ADAS assessment.* Deliverable 2.2.1: Adaptive Integrated Driver-vehicle InterfacE Consortium, final version. Gothenburg, SE: VTEC.

Kiefer, R.J., Flannagan, C.A., & Jerome, C.J. (2006). Time-to-Collision Judgments Under Realistic Driving Conditions. *Human Factors*, *48*, 334-345

Knapp, A., Neumann, M., Brockmann, M., Walz, R., & Winkle, D. (2009). *Code of Practice for the Design and Evaluation of ADAS*. Response 3 - a PReVENT Subprojct, final report. Stuttgart: DCAG.

Lee, J.D. (2008). Fifty years of driving research. *Human Factors*, *50*, 521-528

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., Horst, D., Juch, S., Mattes, S., & Foehl, U. (2005). *Driving performance assessment - methods and metrics*. Deliverable 2.2.5: Adaptive Integrated Driver-vehicle InterfacE Consortium. final version. Gothenburg, SE: VTEC.

Pataki, K., Schulze-Kissing, D., Mahlke, S., & Thüring, M. (2005). Anwendung von Usability-Maßen zur Nutzeneinschätzung von Fahrerassistenzsystemen. In K. Karrer, B. Gauss and C. Steffens (Eds.), *Beiträge zur Mensch-Maschine-Systemtechnik aus Forschung und Praxis* (pp. 211-228). Düsseldorf: Symposion.

Reason, J.T. & Brand, J.J. (1975). *Motion Sickness.* Oxford, UK: Academic Press.

Riedel, A. & Arbinger, R. (1997). *Subjektive und objektive Beurteilung des Fahrverhaltens von PKW*. FAT Schriftenreihe Nr. 139. Frankfurt/M.: Druckerei Henrich.

Schaller, T. (2009). *Stauassistenz - Längs- und Querführung im Bereich niedriger Geschwindigkeit*. PhD thesis, TU München. München: Fakultät für Maschinenwesen.

Spanner-Ulmer, B. (2008). Mensch-Maschine-Kommunikation. Paper presented at the *46. Deutscher Verkehrsgerichtstag - Arbeitskreis VII*, 2008, January. Goslar

Wierwille, W., Tijerina, L., Kiger, S., Rockwell, T., Lauber, E., & Bittner, A. (1996). *Heavy Vehicle Driver Workload Assessment - Task 4: Review of Workload and Related Research.* Final Report Supplement. Washington, D.C., USA: NHTSA.

Success factors for navigational assistance: a complementary ship-shore perspective

Linda de Vries
Chalmers University of Technology, Department of Shipping and Marine Technology
Sweden

Abstract

The maritime domain is under pressure from changing economic, political and environmental factors. Technological advancements facilitate increased monitoring and control from land. By viewing the maritime domain as a complex socio-technical system, the importance of understanding the role of the on board and shore-side operator in maintaining safety and efficiency of navigation becomes apparent, particularly when introducing new technology. This paper looks at the success factors for navigational assistance, as currently performed by maritime pilots and Vessel Traffic Service (VTS) operators, aiming to identify issues worth consideration in future navigational assistance services. One focus group and one combined workshop/focus group were held with three pilots and two VTS operators respectively. The first looked at the prerequisites for successful navigational assistance from the perspective of the pilot. Using a grounded theory-style approach, a proposition was created that the main indicator of success is "no incidents", that success depends on the integration of local knowledge, preparation and foresight into the ship-shore system and that good communication is vital to achieving this. Testing this, the second study considered the role of communication in enabling the VTS operator to support the pilot; it confirmed the results of the first study, emphasising the importance of communication when working both with on board and shore-based pilots.

Introduction

The maritime domain is under pressure from changing economic, political and environmental factors. Modern shipping must deal with an increasing volume and diversity of waterborne transport operating within an ever decreasing navigational space, while simultaneously attempting to curb emissions. Larger vessels are being operated by smaller crews. Shipping routes are being integrated into inter-modal logistics networks. The move towards shipping as part of an integrated transport system brings with it increased demands for information exchange between the vessel and land-based stakeholders and authorities. Various initiatives on a national, European and international level are being put in place to address these challenges, which are pushing the boundaries (Rasmussen, 1997) of the International Maritime Organisation's (IMO) guiding principles of safety and efficiency.

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Advancements in communication and navigation technologies have paved the way for a technical infrastructure in which this information exchange is rapidly becoming reality, allowing for increased centralised monitoring and guidance of vessels. The IMO (2014) have recently finalised a draft Strategic Implementation Plan for e-navigation with the objective to "facilitate a holistic approach to the interaction between shipboard and shore-based users, under an overarching e-navigation architecture" by 2019. The ability to share information between ship and shore also opens up the possibility to introduce new forms of navigational assistance. However, before doing so, it is necessary to understand which services exist today to assist in the navigation of seagoing vessels, how they complement each other, and most importantly, what makes them work (Rochlin, 1999; Johansson & Persson, 2009).

Recent developments in organisational safety such as Resilience Engineering (Hollnagel, 2006) and Safety-II (Hollnagel, 2014) emphasise this same focus on everyday operations, a systemic view in which a successful outcome is created by adapting to the dynamic environment, and safety is often indicated by the absence of incidents (Woods, 2006). Indeed, perspectives from systems engineering dominate the literature on the maritime domain. It is often viewed as a complex socio-technical system (Perrow, 1984; Koester et al., 2007) or a Joint Cognitive System (JCS) (Hollnagel & Woods, 2005), in which the operator interacts with the organisation, technology, physical environment and many other factors, working together to keep the system operating within acceptable parameters and achieve a common goal, in this case the safety and efficiency of navigation. Much of the discourse revolves around control and the link between loss of control and unexpected events. Issues raised are whether safety is improved by centralised (shore-side) or decentralised (on board) control (Perrow, 1984; Weick, 1987; van Westrenen & Praetorius, 2012); the role of both feedback, i.e. input from the environment, and feedforward control, the ability to pre-empt deviations, driven by local knowledge (Hollnagel, 2002; Johansson, 2005; van Westrenen, 1999; Bruno & Lützhöft, 2009); and the importance of achieving tactical (short-term, localised) and strategical (longer term, system-wide) control (Praetorius, 2014; Praetorius & Hollnagel, 2014), although this is often not achieved in practice (Hollnagel & Woods, 2005). This paper, however, attempts to step back and describe some preliminary investigative studies into the success factors for various forms of existing navigational assistance from the perspective of the operator, with a starting point in practice, rather than theory.

Overview of existing navigational assistance services

"Navigational assistance" is an overarching term encompassing several forms of service which aim to assist the ship's captain, known as the "master", with the safe navigation of their vessel in areas where this is deemed necessary. It will be used in this paper to include pilotage services, both on board and shore-based, and navigational assistance as performed by Vessel Traffic Services (VTS). It should be noted that although this inclusive term is utilised by the author, it is not necessarily used by the practitioners.

Pilotage
Pilotage has a long and well-established history, stretching back at least 4,000 years (IMPA, 2014). Pilotage can be defined as "to guide vessels into or out of port safely - or wherever navigation may be considered hazardous, particularly when a shipmaster is unfamiliar with the area" (IMO, 2014), comprising "activities related to navigation and ship handling in which the pilot acts as an advisor to the master of the ship" (IALA, 2012a). Pilotage is generally conducted on board the vessel (Hadley, 1999; van Westrenen, 1999; Grundevik & Wilske, 2007) but, in some areas and in certain, often weather-related, circumstances, remote pilotage i.e. from "a position other than aboard the vessel concerned" (Hadley, 1999; EMPA, 2014) may also be conducted.

Vessel Traffic Services
Vessel Traffic Services (VTS) is a shore-based service, established to "improve the safety and efficiency of vessel traffic and to protect the environment", offering one or more of three levels of service: information service (INS), navigational assistance service (NAS) or traffic organisation service (TOS) (IMO, 1967). NAS, a service "to assist on-board navigational decision-making and to monitor its effects", is usually requested by the vessel (van Westrenen & Praetorius, 2012) or given when observed to be necessary by the VTS (IALA, 2012b). The vessel is recommended, but not obligated, to follow this advice (IMO, 1967). In practice, there is no sharp distinction between INS, NAS and TOS (Praetorius, 2014), and all may be seen as, directly or indirectly, assisting in the safety of navigation.

Responsibility for safety of navigation
Although both pilots and VTS operators may provide advice on navigational matters, responsibility for safety of navigation remains at all times with the master of the vessel (STCW, 1995/2010; COLREGS, 1972). The VTS operator or pilot do not relieve the master of this responsibility (IMO, 1967; IALA, 2012a).

Method

The general approach can be described as grounded theory-inspired, taking elements of grounded theory as developed by Glaser and Strauss (1967) (also Charmaz, 2000). A variety of methods and data sources were used in order to create and develop a general "proposition" concerning the success factors for navigational assistance. This was treated as a substantive theory (Denscombe, 2010), a localised, empirical theory, or a general statement about the phenomenon to be subsequently confirmed, refuted or amended, and was indeed used in this way throughout the remainder of the studies.

The process did not strictly follow the step-by-step procedure as originally described (Glaser & Strauss, 1967; Czarniawska, 2014), being more opportunistic and pragmatic in nature. As one aim of the studies was to feed the results back into the maritime community, it was considered important that the outcome be recognisable and relevant to practitioners. Therefore an approach with links to pragmatist thinking was used (Locke, 2001). Data collection was mainly done through a focus group (Corbin & Strauss, 2008; Stanton et al., 2006) and a workshop with expert practitioners. Field observations and informal conversations with various

stakeholders were also used to complement the data, utilizing a series of "double-back steps" (Glaser, 1978) to continuously refine the emerging results. The diversity of methods, data sources and materials was considered useful in highlighting different aspects of the topic (Glaser, 1978: Strauss, 1987). As the method and the results are very much intertwined in this approach, a description of how the studies were conducted, progression was made throughout and results were developed iteratively will be shown in this section; the actual results will be included in the following section.

Focus group with deep sea pilots

The studies commenced with a focus group looking at the success factors for navigational assistance from the perspective of the maritime pilot, more specifically, the deep sea pilot. The focus group consisted of three deep sea pilots operating in the Baltic Sea and Kattegatt/Skagerack areas. The pilots had similar backgrounds but varying levels of professional experience and length of service. The participants were given one open-ended question which was then discussed in detail with very little intervention from the moderators. They themselves described in very clear terms what they considered the success factors, and in particular how success is measured (see results). As they were emphatic on this point, their phrasing was retained and, by using constant comparison throughout the analysis, its centrality was confirmed. Likewise, the participants themselves identified the relationships between various types of information, which would become the categories and themes of the analysis, already during the focus group, and the importance of communication of this information between ship and shore. Thus, much of the analysis took the form of a cross-check on the data, rather than an analysis per se; it merely confirmed the relationships between the factors already identified by the participants.

As all the participants and researchers present were either native speakers, with the exception of the author who has a good working knowledge of the language, the focus group was held in Swedish. Transcriptions were made in the original language and loosely translated by the author. The transcriptions were coded and analysed iteratively using an inductive approach. Comparison was conducted throughout with photographs of the participants' brainstorming on the whiteboard and the authors' own notes. As codes and categories emerged, the wording was kept as close to the original as possible. In most cases, a direct translation into English was considered sufficiently accurate. Open coding produced a large number of categories which were then, by a process of axial coding, interlinked and consolidated into themes and related to a central concept (Strauss & Corbin, 1990) from which the proposition was developed. A table and corresponding diagram showing the relationships between categories, topics and main concept was generated and from this the proposition was formulated (Figure 1).

A very preliminary version of the proposition was presented in text and diagram form to pilots and VTS operators at a project meeting and received positive feedback. Informal conversations revealed support for the proposition from the point of harbour and coastal pilots ("difference minimisation", Glaser & Strauss, 1967) as

well as deep sea pilots. One of the participants in the focus group also confirmed that this was a true representation of their discussion.

Workshop with Vessel Traffic Service (VTS) Operators

Having considered the success factors from the perspective of the on board pilot, and having received confirmation, albeit on a limited scale, from a wider community of pilots and VTS operators, the phenomenon was investigated further by looking into how communication between ship and shore contributes to successful operations from the point of view of the shore-side operator, the VTS operator ("difference maximisation", Glaser & Strauss, 1967). Initially the second data collection was intended to be another focus group mirroring the first but from the shore-side perspective. It was to form part of a larger expert workshop looking at everyday operations in the VTS domain. However, due to availability of participants, only two were able to remain for the part of the workshop which is described in this study and the format was thus revised.

The participants, experienced VTS operators working in two large European ports, were first asked individually to describe their VTS areas by drawing a map on the whiteboard, and then describe the process of communication between the VTS and pilots by annotating on the map. They were then asked, in a group interview style, to discuss what makes for successful communication between the VTS and the pilots, what can be improved. Since both on board and shore-based pilotage are available in their areas, they were asked how the communication changes in the case of the pilot being shore-based as opposed to on board the vessel. Photographs of the maps and diagrams drawn by the participants were taken and the discussions were voice recorded. The language used was English. Once again, transcriptions were made of the recordings and a loose open coding conducted. However, instead of developing categories from the wording of the discussions, the categories developed in the analysis of the first focus group were used to sort the data. These were deemed to be mainly sufficient, though a couple of new categories were added. Throughout this sorting process, the participants' maps and drawings were continuously referred to, as were the table and diagram of results and the proposition from the previous study. The table and diagram were then annotated to show how the findings from the second study confirmed or refuted those of the first, and to show any new data which had emerged. As certain aspects were identified as being of greater importance during the second study, these were also highlighted.

Field observations and futher informal conversations with practitioners

Following the workshop with the VTS operators, further low key data collection was conducted over a period of several months to observe how the factors identified by the respondents manifest themselves in practice. This included the shadowing of a harbour pilot in their daily work: receiving the pilot booking from the VTS, transfer with the pilot boat to the vessel, boarding the vessel at the pilot boarding point, connecting the tugs and berthing the vessel in the harbour. Several informal conversations and observations have also been held with pilots and VTS operators, both in their operational environment and in training situations, such as the pilot station, VTS centre, classroom and VTS simulators.

Results

Focus group with deep sea pilots

The main findings of the first focus group with the deep sea pilots were that:

The main indicator of successful navigational assistance is "no incidents". This is dependent on (i) the pilot as the link in the chain of communication between the vessel and the VTS and (ii) the integration of information based on local knowledge, preparation and foresight.

A surprising finding (at least for the author) was the respondents' unanimous insistence on "no incidents" as the trademark of a successful assistance, rather than, as might have been expected, safety. Incidentally, when the moderator tried to categorise "no incidents" as "safety" during the discussion, the respondents interjected with a comment that "What is safety? We can't measure it, but what we can see is that nothing went wrong." The above formulation emerged from the grounded theory-style analysis of the data. This was then used as the proposition to be further investigated in the remaining studies. A visual representation is shown below.

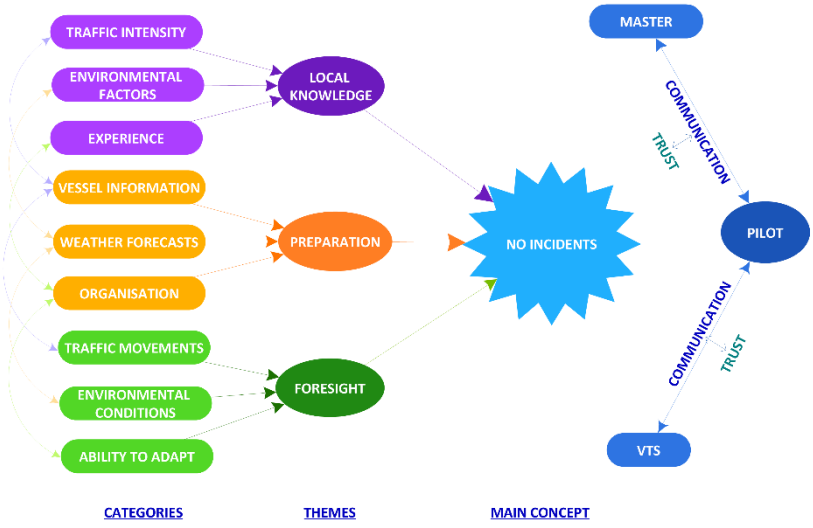


Figure 1. Success factors from the perspective of the pilot

Local knowledge is made up of information about (a) the traffic intensity, such as the types, sizes, speeds and schedules of vessels operating in the area and which routes they tend to take; (b) environmental factors such as weather patterns, water depth, currents, water level etc.; (c) experience, not just in terms of training and time as a captain or pilot, but also regarding the interpretation of information from the environment, other vessels, technology, the vessel crew and the VTS.

Preparation is mainly concerned with (a) vessel and traffic information, both in terms of receiving the details of the vessel to be piloted, size, type, crew, shiphandling characteristics, destination and estimated time of arrival (ETA), but also the expected traffic situation and intensity; (b) weather forecasts, including predicted wind, visibility, waves, currents and water level; (c) organisation, including factors such as scheduling, possibilities for rest periods, travel, handovers between pilots, as well as the ability of the pilot to receive and assimilate weather and vessel information and create a plan for the voyage.

Foresight is built on a combination of (a) vessel and traffic movements, both the shiphandling of the vessel being piloted in the current traffic situation and weather conditions, and the interaction with other vessels and VTS in the area; (b) environmental conditions, the effect they are having on the vessel and traffic movements; (c) ability to adapt to the vessel and its crew, other traffic, weather etc. in order to avoid incidents and keep the appointed ETA.

The pilots perceived their role as integrating the aforementioned information and being the link in the chain of communication between the vessel crew, particularly the master, and the shore-side VTS operators. While they acknowledged that the level of communication and cooperation between parties may depend on culture, nationality and role of the different parties, and is not always optimal (see also TSBC, 1995), they emphasised that communication is usually successful because of the inherent trust in the role of the pilot (see Meyerson et al., 1996; Bruno & Lützhöft, 2010); they are welcomed on board and seen as part of the bridge team, bringing their local knowledge, preparation and foresight to the situation and bridging the language gap between the ship and shore (also noted in van Westrenen 1999, 2011; van Westrenen & Praetorius, 2012) .

Workshop with Vessel Traffic Service (VTS) Operators

The second study, the workshop with VTS operators, confirmed the results of the first. Although the focus was on communication, all the categories identified as success factors in the first study were mentioned as being instrumental by the VTS operators, with a particular emphasis on preparation. The main findings were thus that:

Success is dependent on good communication between the VTS, pilot and vessel, being especially critical in the preparation phase.

Additionally, a number of new issues were highlighted: the importance of communication between the pilot and tugs and fishing vessels; the role of the pilot as the interpreter between the vessel, where English will usually be the language used with the crew, and the VTS and tugs, where the communication may often be in the local language; that changes in routine and particularly in the co-location of the VTS and pilot services may have major impact on the communication between the parties (see also Praetorius, 2014). This was considered true regardless of whether the pilot is on board or shore-based. The findings from the second study were used to annotate and modify the results diagram as shown in figure 2.

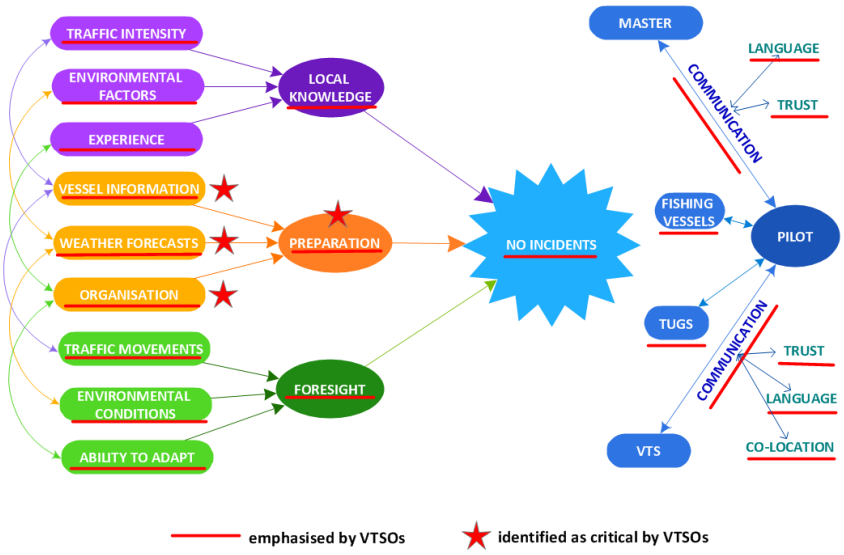


Figure 2. Success factors from the perpective of the pilot and VTS operator

Field observations and futher informal conversations with practitioners

During the field observations of the pilot at work and other informal conversations with pilots and VTS operators, the findings from both the studies with the deep sea pilots and the VTS operators were again confirmed. In particular, the inherent "status" of the pilot as the local navigation expert (as noted in van Westrenen, 1999; Darbra et al., 2007) as soon as they step on board the vessel, and the ability to quickly build a relationship of trust (see also Meyerson et al., 1996; Bruno & Lützhöft, 2010) with the master and crew using verbal and non-verbal communication (see Flin et al., 2004) were noted. The role of the pilot as interpreter between the tugs and vessel was also apparent. Also noticeable was the proactive nature of both the pilot and VTSO operator at work; continuously scanning the information available to them, weighing up options, planning the next steps and adapting the language and content of their communication to effect the required response from the vessel crew.

Discussion

The number of respondents in both the focus group and the workshop was unfortunately very low and not representative of the population as a whole. It was established that the participants of the first focus group also had considerable experience operating within harbour and coastal pilotage areas and so were qualified to represent not only deep sea pilots, but pilots in general. Still, anchoring in a wider population of pilots and VTS operators is needed. Also, when considering the interaction between the pilot, VTS and vessel, it is of course necessary to consider

the perspective of the master and crew of the vessel being assisted. This remains to be done in the near future. The author believes that the results are nevertheless indicative and that further studies following this cumulative, flexible approach approach (Corbin & Strauss, 2008) will confirm this.

From both the ship and shore-side perspectives, two main points were emphasised throughout: that success is dependent on *communication* and *integration of information*. The proposition developed during the first study with the deep sea pilots was thus confirmed. This rather simplistic proposition, easily recognisable to practitioners, almost to the point of being too obvious (consider Czarniawska's (2014) comment on grounded theory being "nothing more than the common sense of fieldwork"), nonetheless disguises the complexity of the services provided by pilots and VTS operators within the maritime domain. It focuses on the ability of the human operator within the system to integrate and communicate, without going in detail into the vast range of sources of information being integrated and communicated; the various means and timescales within which this is being realised; and the dynamic and unpredictable nature of many of the elements and the interaction between them. Only by observing them at work can this complexity truly be appreciated.

In addition to building on this empirical approach, a parallel examination of navigational assistance from a theoretical perspective may provide additional insight and give weight to the findings so far established. It became apparent during the analysis of the second study that, while the practitioners, particularly the pilots, identified and categorised their work as the integration of local knowledge, preparation and foresight, within each of these topics, another pattern may be identified; each is based upon information regarding (i) vessels and traffic, (ii) weather and physical environment and (iii) the skills and characteristics of the operator. In other words, the ship-shore interaction may be seen as the human, technical and environmental elements of a complex socio-technical system (Perrow, 1984) or joint cognitive system (Hollnagel & Woods, 2005).

Furthermore, the distinction made, consciously or unconsciously, by the operators can be regarded as relating to different but interrelated aspects of time: (i) local knowledge, about traffic, environment and other factors, is built over a long time period, but once established is fairly constant; (ii) preparation is concerned with the hours or day before the navigational assistance takes place; (iii) foresight deals with the present and near future. Integration and communication of information on all three time scales is necessary to ensure success, creating the preconditions for what may be described as strategic or tactical control in a resilient system (Praetorius & Hollnagel, 2014). Although the participants do not speak in terms of control, talking instead of communicating information between ship and shore, aspects of both centralised control, e.g. the VTS coordinating pilot boarding, and decentralised control, such as the pilot directing the tugs, may be seen.

More problematic is the apparent paradox that, while the official goal of both on board and shore-side navigational assistance is the safety and efficiency of navigation (IMO, 1967; 1969), in practice a successful outcome is "no incidents". Safety is seen as a dynamic non-event (Weick, 1987). Indeed, according to the

practitioners in these studies, it is unmeasurable. While they can identify the necessary ingredients to create success, and observe these in their daily work, they cannot identify success itself other than as the absence of failure. They are also divided on the extent to which success is partially attributable to chance. This same paradox is discussed in some detail within the field of resilience engineering (Hollnagel, 2014), and appears to be one of the major challenges to be met in order for organisations concerned with safety in dynamic conditions to change their focus towards success in everyday operations. This also has implications for the design of future navigational assistance services, if they are to achieve a positive measure of safety, rather than be characterised by a lack of failure.

Conclusions

The main conclusions to be drawn from this simple preliminary study are that, from the perspective of the on board and shore-side operator:

The main indicator of successful navigational assistance is "no incidents". This is dependent on (i) the pilot as the link in the chain of communication between the vessel and the VTS and (ii) the integration of information based on local knowledge, preparation and foresight.

It is hoped that further investigations, both in terms of additional empirical data collection and an examination of the phenomenon from a theoretical perspective, will contribute to a set of preconditions for successful navigational assistance which should be considered in the development of future maritime communication infrastructures and e-navigation services.

References

Bruno, K. & Lützhöft, M. (2009). Shore-based pilotage: Pilot or autopilot? Piloting as a control problem. *Journal of Navigation, 62*, 427-437.

Bruno, K. & Lützhöft, M. (2010). Virtually Being There: Human Aspects of Shore-based Ship Assistance. *WMU Journal of Maritime Affairs, 9* (1), 81–92.

Charmaz, K. (2000). Grounded theory: Objectivist and constructivist methods. In N Denzin and Y. Lincoln (Eds.), *Handbook of qualitative research*, 2nd ed. (pp. 609-635). Thousand Oaks, CA: Sage.

Charmaz, K. (2006/2014). *Constructing grounded theory*, 2nd edition. London: Sage.

Corbin, J.M. & Strauss, A.L. (2008). *Basics of Qualitative Research: Techniques and procedures for developing grounded theory*. Thousand Oaks, CA: Sage.

Czarniawska, B. (2014). *Social science research: from field to desk.* Lund: Studentlitteratur.

Darbra R.M., Crawford J.F.E., Haley C.W., & Morrison R.J. (2007) Safety culture and hazard risk perception of Australian and New Zealand maritime pilots. *Mar Policy* 31, 736–745. doi:10.1016/j.marpol.2007.02.004

Flin, R., & Burns, C. (2004). The role of trust in safety management. *Human Factors and Aerospace Safety, 4*: 277–87.

Glaser, B. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory.* Mill Valley, CA: Sociology Press.

Glaser, B. & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.

Grundevik, P. & Wilske, E. (2007). Uppdrag avseende ny teknik för lotsning. *SSPA Sweden AB, rapport 2007 4449-1.*

Hadley, M. (1999). Issues in Remote Pilotage. *Journal of Navigation*, *52*, 1–10.

Hollnagel, E. (2002). *Cognition as Control: A Pragmatic Approach to the Modelling of Joint Cognitive Systems.* http://www.ida.liu.se/yeriho/images/IEEE_SMC_Cognition_as_control.pdf

Hollnagel, E., 2014. *Safety-I and Safety-II: The past and future of safety management.* Farnham: Ashgate.

Hollnagel, E. (2006) Resilience: the challenge of the unstable. In E. Hollnagel, D. Woods, and N. Leveson (Eds.), *Resilience Engineering: Precepts And Concepts* (pp. 21-34). Abingdon: Ashgate Publishing Group.

Hollnagel, E. & Woods, D. (2005). *Joint Cognitive Systems: An Introduction to Cognitive Systems Engineering*. London: CRC Press.

IALA (2012a). *Pilotage Authority Forum (PAF) Report on Best Practice for Competent Pilotage Authorities.* Edition 1.1. Saint Germain en Laye, France: International Association of Marine Aids to Navigation and Lighthouse Authorities

IALA (2012b). *IALA—Vessel Traffic Services Manual*, 6 edn. Saint Germain en Laye, France: International Association of Marine Aids to Navigation and Lighthouse Authorities,

IMO (1967). *IMO Resolution A.857(20) Guidelines for Vessel Traffic Services*. London: International Maritime Organisation.

IMO (1968). *IMO Resolution A.159(ES.IV) Recommendation on Pilotage*. London: International Maritime Organisation.

IMO (2014) *Proceedings of Sub-Committee on Navigation, Communications and Search and Rescue (NCSR),* http://www.imo.org/MediaCentre/MeetingSummaries/NAV/Pages/NCSR-1st-Session.aspx

IMPA (2014). *IMPA on Pilotage*. Livingstone: Witherby Seamanship International.

Johansson, B. (2005). *Joint control in dynamic situations*. Doctoral dissertation, Linköping University, Linköping, Sweden. (Linköping Studies in Science and Technology: 972).

Johansson, B., and Persson, P.-A. (2009). Reduced uncertainty through human communication in complex environments. *Cognition, Technology & Work, 11*, 205–214.

Koester, T., Anderson, M. & Steenberg, C. (2007). *Decision Support for Navigation*. FORCE Technology, Draft Report DMI 107-27358.

Locke, K. (2001). *Grounded theory in management research*. London: Sage.

Martin, P. and Turner, B. (1986). Grounded theory and organisational research. *The Journal of Applied Behavioural Science*, *22*, 141-157.

Meyerson, D., Weick, K. and Kramer, R. (1996). Swift trust and temporary groups. In R.M. Kramer and T.R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp.166–195). Thousand Oaks: SAGE Publications.

Perrow, C. (1984). *Normal Accidents*. Princeton, New Jersey: Princeton University Press.

Praetorius, G. (2014). *Vessel Traffic Service (VTS): a maritime information service or traffic control system?*. PhD thesis, Gothenburg: Chalmers University of Technology. ISBN/ISSN: 978-91-7597-048-6.

Praetorius, G. & Hollnagel, E. (2014). Control and resilience within the maritime traffic management domain. *Journal of Cognitive Engineering and Decision Making December 2014 8:* 303-317, DOI:10.1177/1555343414560022.

Rasmussen, J. (1997). Risk management in a dynamic society: a modelling problem. *Safety Science Vol. 27, No. 2/3*, 183-213.

Rochlin, G.I. (1999). Safe operation as a social construct. *Ergonomics, 42*, 1549–1560.

Strauss, A. (1987). *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.

Strauss, A.L. and Corbin, J. (1990). *Basics of qualitative research: grounded theory procedures and techniques*. London: Sage.

van Westrenen, F. (1999). *The Maritime Pilot at Work. The Evaluation and Use of a Time-to-Boundary Model of Mental Workload in Human-Machine Systems.* Doctoral dissertation, Delft University of Technology, Delft, the Netherlands.

van Westrenen, F. (2011). Cognitive work analysis and the design of user interfaces. *Cognition, Technology & Work, 13*, 31–42. DOI 10.1007/s10111-010-0153-4.

van Westrenen, F. & Praetorius, G. (2012). Maritime traffic management: a need for central coordination?. *Cognition, Technology & Work, 16*, 59-70 DOI 10.1007/s10111-012-0244-5.

Transportation Safety Board of Canada (1995). *A Safety Study of the Operational Relationship Between Ship Masters/Watchkeeping Officers and Marine Pilots.* Report number SM9501.

Weick, K. E. (1987). Organisational culture as as source of high reliability. *California Management Review, 16(3),* 571-593.

Woods, D. (2006). Essential Characteristics for Resilience. In E. Hollnagel, D. Woods, and N. Leveson (Eds.), *Resilience Engineering: Precepts And Concepts* (pp. 21-34). Abingdon: Ashgate Publishing Group.

Can weak-resilience-signals (WRS) reveal obstacles compromising (rail-)system resilience?

Willy Siegel[a] *& Jan Maarten Schraagen*[a,b]
[a]*University of Twente, the Netherlands*
[b]*TNO Earth, Life, and Social Sciences, the Netherlands*

Abstract

Analysis of accidents in socio-technical systems frequently reveals unnoticed obstacles, which have grown to become the main cause of incubation and surprise at failure (Dekker, 2011). Thus far, it has proven to be a challenge to identify those unnoticed obstacles upfront among the tremendous number of events occurring during normal operations. In this article, we describe the usage of weak resilience signals (WRS) (Siegel & Schraagen, 2014), at a rail control post, to reveal obstacles compromising the resilience state of the system. Resilience is defined as the ability of a complex socio-technical system to cope with unexpected and unforeseen disruptions (Hollnagel, Woods, & Leveson, 2006). The WRSs, developed and presented around three system boundaries: safety, performance and workload, are used to stimulate a state of mindfulness (Weick & Sutcliffe, 2007) revealing unnoticed obstacles. An observational study is proposed to verify exposure of obstacles and their impact on rail-system resilience. The WRS and its stimulus to rail traffic controllers are expected to contribute to a higher rail operation reliability.

Introduction

Accident analyses of socio-technical systems expose unnoticed disturbances which are a component in the process towards failure (Hall, 2003; Stanton & Walker, 2011). These disturbances are either not observed or ignored throughout the complex process of the system. This is not surprising since many disturbances occur continuously and do not evolve into an accident. Some disturbances are identified with a potential to evolve into an accident, but are ignored due to the culture of the organization (Vaughan, 1997, 2002). Weick and Sutcliffe (2007) propose high-reliability-organization principles influencing the culture of organizations to deal with the unexpected. They introduce the term ‘mindfulness’, split into the phases anticipation and containment, to work out the principles. The three principles of anticipation are: 1) preoccupation with failure; 2) reluctance to simplify; and 3) sensitivity to operations. The two additional principles of containment are: 4) commitment to resilience; and 5) deference to expertise. In previous research, we have developed weak-resilience-signals (WRS) to identify disturbances to the resilience state of a rail-system (Siegel & Schraagen, 2014). The WRSs are signals around the boundaries: safety, performance and workload, on a high aggregation level needing further analysis to understand the root causes. We described a method

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

to measure workload WRS and applied it at a rail control post. Analysis of the workload WRS identified has revealed a disturbance which we call an obstacle. The obstacle identified influenced the resilience of the system. Our analysis showed that the obstacle attracted resources and attention, which may influence the spare capacity needed when a disruption occurs. Although the WRS measurement has a clear methodology, the obstacle identification has not and was a result of ad hoc analysis. This was sufficient to quantify a WRS, since it proved the ability of a WRS to reveal an obstacle, but left a gap concerning the methodology of obstacle identification. The aim of this article is to fill this gap by describing a process to reveal obstacles systematically using WRSs as the carrier of mindfulness.

Process to reveal obstacles with help of WRSs

The first principle of Mindfulness, defined by Weick & Sutcliffe (2007) as "a rich awareness of discriminatory detail", is preoccupation with failure. They suggest four questions to deal with this principle which will cause "actively searching for weak signals that the system is acting in unexpected ways" (Weick & Sutcliffe, 2007, p. 151):

> 1) What needs to go wrong?; 2) What could go wrong?; 3) How could things go wrong?; 4) What things have gone wrong?

The focus is here on "wrong", occurring repetitively in all questions, while a central concept of resilience is the focus on what goes right (Hollnagel, 2009). We suggest to seek beyond failure to enrich Mindfulness using weak-resilience-signals (WRSs).

The Mindfulness phase of anticipation is suitable to enrich with signals to anticipate on like the WRS, while the containment phase is about the way of acting and has no direct relation with signals. Therefore, we only adjust the three Mindfulness principles of anticipation, marked with underlined text, to focus on the WRS and are phrased as follows:

1. Preoccupation with *WRS in addition to* failure;
2. Reluctance to simplify *WRS interpretations*;
3. Sensitivity to operations *by being aware of WRS*.

The *preoccupations with WRS, in addition to failure*, can be achieved through after-shift-review discussion of a rail traffic control team guided by questions they have to answer. The team has to go through a process of analysing the WRS based upon the activities occurring throughout its shift. For doing that, they need *sensitivity to the operation* and keep in mind operational facts to be used at the review. During the review they *should not simplify the reasoning* of the WRS but stimulate each other for deep reasoning and search for underlying reasons and conditions causing the WRS beyond their own responsibility. Once rail traffic controllers have understood these conditions, they have to discuss whether they can reoccur as an obstacle to interfere with future operations. Finally, they have to discuss how they can anticipate these obstacles.

A set of after-shift-review questions will help the team to direct its discussion:

- Which conditions have made the WRS possible? Search deep and beyond your responsibility.
- Are (some of) these conditions obstacles that may reoccur?
- Which actions can be taken, on different levels of the system, to anticipate these obstacles?

The first question causes the team to think in terms of conditions, rather than obstacles. They should not simplify these conditions on their relative small span of control, but search beyond the responsibility of the individual and the team. When understanding the conditions, they can progress to the second question dealing with potential reoccurrence. Reoccurrence is an important attribute of an obstacle in addition to the ability to respond to the occurrence. This ability is the core of the third question, dealing with anticipation. Different levels of the system can anticipate. Anticipation is possible on the level of the individual and the team. In this case, the team can agree on future actions to take. However, some anticipatory action can only be taken on higher levels, like the whole Post, the national control centre, the company or even on the national political level.

To illustrate the above, we will take a workload WRS identified by Siegel and Schraagen (2014). This workload WRS presents a situation of a rail controller being occupied during the morning shift by continuous ad hoc shunting activities, rating his workload the whole morning much above the standard low workload. The standard low workload enables him to peak and react adequately when an unexpected disturbance occurs. The continuous ad hoc shunting activities may undermine his ability to react appropriately. A discussion of the team about this workload WRS, with help of the above review questions, can result in the following. The team identifies the condition that small train companies using the rail infrastructure are having difficulties to manage their equipment and react on the spot without planning shunting movements ahead. This situation is reoccurring and can be seen as an obstacle, since it occupies the spare capacity needed during calamities, causing a reduction in resilience. Anticipation on this obstacle is possible on different levels. The individual rail controller can either request his counter party to plan his activities ahead or refuse accepting the shunting order. The team can reorganise its activities to unload the specific rail controller to manage its capacity. The Post, being the management unit of the teams, can add resources to the team to bring the workload to the standard level or approach the local management of the train companies to search for a solution. This obstacle can also be dealt with on a national level, which goes beyond the direct influence of the team, but could be addressed by the Post management.

The proposed process needs to be verified and prove its ability to expose obstacles, compromising system resilience. In the next section, we describe the design of an observational study at a rail control post to verify the process in a socio-technical rail-system.

Observational study design at a rail control post

The main effect to verify the proposed process is its influence on the resilience state of the system. Hollnagel (2009) states that resilience implies four essential system

capabilities, also called the four cornerstones of resilience: anticipating, responding, monitoring, and learning. The proposed process of using WRSs at after-shift-review s aims to improve 1) the *learning* of team performance throughout their shift and 2) the *anticipation* on the obstacles identified. In that sense, the verification should focus on learning and anticipation to prove the influence on the system resilience. However, this does not imply the resilience compromise of the obstacles identified. Analysis of scenarios describing the obstacle occurrence, with help of all four cornerstones with emphasis on *responding* and *monitoring*, can indicate the resilience impact of the obstacle itself. Another aspect to verify is the influence of the WRS itself on the whole process. In other words, what would be the result of conducting an after-shift-review of the events, without presenting the WRSs? We will address these aspects in the study design after describing the setting at the rail control post, where the observation takes place.

The setting is a rail control post responsible for an area with rail stations split up into two main rail corridors: south-north, called corridor North, and west-east, called corridor East. Each of the corridors has workstations for rail controllers working in three shifts operating the control post 24 hours a day. Corridor North has 4 workstations, corridor East has 3 workstations, and one workstation at the post is used only during calamities and can be added to each corridor. At the Post, approximately 70 rail controllers are authorized to work at one, more, or all of the workstations. During a trial period of one week, the morning shift of corridor East will conduct an after-shift-review discussion for an hour. The first half hour will concentrate on the occurrences of the day and the second half hour on WRSs as described in the previous section. Corridor North and the other shifts will not conduct an review. The review will be led by a team-leader, who is not a rail-controller, and observed by a researcher. The researcher will take notes on the discussion and focus on the difference in the two half hours and on the reasoning trace of the obstacles. After the review, the researcher will interview each team member of corridor East, and of corridor North and of the next shift of corridor East as reference.

The researchers will seek for evidence through interviews on the hypothesis that: 1) the resilience of the morning shift of corridor East grows and 2) the resilience has grown due to the review discussion on WRSs. The first hypothesis will be tested by: 1) an observed growth of learning and anticipation plans and 2) identification of obstacle scenarios influencing the four cornerstones. The findings will be corroborated through interviews with the target and reference teams. The second hypothesis will be tested through the difference between the first and second half hour of the review as well as with interviews with the different teams.

Summary and discussion

We combine in this article two theories, high-reliability-organisations and weak resilience signals (WRS). High-reliability-organisations underpin their qualities with the assumption that first, it is possible to identify and anticipate potential failure scenarios, and second, it is possible to spot errors when they occur and identify a timely and appropriate course of action in real time to avert catastrophic consequences (Lekka, 2011). Weak resilience signals originate by obstacles which

compromise system resilience but lack a systematic organisational process identifying the obstacles and ensuring the anticipation to prevent their incubation (Siegel & Schraagen, 2014). The two theories seem complementary, where the first concentrates on the organisation and its processes, the second focuses on visualization of cues, which have not been spotted or cannot be seen. However, evidence is needed that in reality they will strengthen each other. We proposed an observational study in a rail operations control room where high-reliability-organisation principles are using weak resilience signals. The study will verify and challenge the hypothesis that weak-resilience-signals can reveal obstacles compromising rail-system resilience. A positive outcome is expected to contribute to a higher rail operation reliability.

Acknowledgement

This research was conducted within the RAILROAD project and is supported by ProRail and the Netherlands organization for scientific research (NWO) (under grant 438-12-306).

References

Dekker, S. (2011). *Drift into failure - from hunting broken components to understanding complex systems*. Farnham, Surrey: Ashgate Publishing Limited.

Hall, J.L. (2003). Columbia and Challenger: organizational failure at NASA. *Space Policy*, *19*, 239–247. doi:10.1016/j.spacepol.2003.08.013

Hollnagel, E. (2009). The four cornerstones of resilience engineering. In C. P. Nemeth, E. Hollnagel, & S. Dekker (Eds.), *Resilience Engineering Perspectives. Volume 2: Preparation and restoration* (pp. 117–134). Surrey: Ashgate Publishing Limited.

Hollnagel, E., Woods, D.D., & Leveson, N. (Eds.). (2006). *Resilience engineering: concepts and percepts*. Hampshire: Ashgate Publishing Limited.

Lekka, C. (2011). High reliability organisations: A review of the Literature. *Health and Safety Executive*. Retrieved from http://www.hse.gov.uk/research/rrpdf/rr899.pdf

Siegel, A.W., & Schraagen, J.M.C. (2014). Measuring workload weak-resilience-signals (WRS) at a rail control post. IIE Transactions on Occupational Ergonomics and Human Factors2(3-4), 179–193. doi:10.1080/21577323.2014.958632

Stanton, N.A., & Walker, G.H. (2011). Exploring the psychological factors involved in the Ladbroke Grove rail accident. *Accident, Analysis and Prevention*, *43*(3), 1117–27. doi:10.1016/j.aap.2010.12.020

Vaughan, D. (1997). The trickle-down effect: policy decisions, risky work, and the Challenger tragedy. *California Management Review*, *39*(2), 80–102.

Vaughan, D. (2002). Signals and interpretive work: The role of culture in a theory of practical action. In K. A. Cerulo (Ed.), *Culture in mind: Toward a sociology of culture and cognition* (pp. 28–54). New York: Routledge.

Weick, K.E., & Sutcliffe, K.M. (2007). *Managing the unexpected: Resilient performance in an age of uncertainty*, *2nd edition*. John Wiley & Sons, Inc.

Introducing electric vehicle-based mobility solutions – impact of user expectations to long and short term usage

André Dettmann[1], Dorothea Langer[2] ,Angelika C. Bullinger[1], & Josef F. Krems[2]
[1]Chair for Ergonomics and Innovation Management
[2]Chair for Cognitive and Engineering Psychology
Chemnitz University of Technology
Germany

Abstract

When introducing innovative technologies, it is crucial that they comply with users' needs or help fulfilling certain tasks. Hence, knowing users' needs and expectations allows developing an innovative technology that motivates high usage and acceptance. The paper presents results of a field study investigating the introduction of a mobility service based on electric vehicles with 120 participants. While introducing the service, user needs and expectations were examined using a semi-structured guided telephone interview and two questionnaires. After launching the service, the actual usage of the system by the users is tracked by collecting system data and conducting ongoing questionnaires. Results of the empirical study show that users' expectations split primarily into two groups. One group perceives the introduced mobility service as a flexible and quick solution to optimize their mobility needs. The second, technology driven group is highly interested in the electric vehicles. System data shows how both groups perform over time to answer questions if usage meets the expectations of both groups and how they influence their overall short and long term acceptance. Results can be integrated in other services/ systems to better address users' needs.

Introduction

Usually people use their private cars with regular combustion engines for short distances. In Germany, the average usage of private cars is about one hour with just one passenger. Furthermore, the average distance driven within urban areas is less than 45 kilometres (Mobilität in Deutschland, 2008). Given these figures and the upcoming scarcity of fossil fuel, the German government demands innovative mobility concepts and changing mobility behaviour. It announced the aim to increase the number of electric vehicles on a large scale by 2020 (Die Bundesregierung, 2009). Electric vehicles use electric power instead of fossil fuels and are able to manage most of the daily transportation tasks within urban areas.

One approach to realize increased usage of electric vehicles is the implementation of mobility-on-demand systems that are characterized by the sequential use of (electric)

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

cars within a service area by different people. Companies, public institutions and authorities which are divided into more than one location are examples for such systems. The implementation could also be enriched by offering (electric) bicycles or special public transport services to provide the users a multimodal mobility solution. The main challenge of multimodal mobility-on-demand systems, i.e. systems which combine public transportation with rental cars and even bicycles, is user acceptance and adoption. When introducing such innovative technologies, it is crucial that they comply with certain user needs or help fulfilling certain tasks. Beliefs that the innovation can meet those requirements can be described as positive user expectations. Knowing those user needs and expectations allows developing a system that motivates high usage and acceptance.

The paper draws on a real-life multidisciplinary research project with both public and privately owned mobility companies as partners. In our research, we used an explorative mixed-method approach, based on qualitative data analysis triangulated with gathered system data to answer the question if user expectations have an impact to long and short term usage and if so, what implications for a mobility-on-demand systems can then be extracted. Furthermore, we want to answer the question, if expectations are a suitable measure for short- and long-term usage predictions. The research group investigates users' needs and expectations using a semi-structured guided telephone interview and two questionnaires. In the remainder of this paper, the next chapter presents the research field with a short summary of prior studies and afterwards the results of our recent studies.

The research field and prior studies

The described mobility solution was implemented at a medium sized German university in 2012. The university has four sites and approximately 4000 employees. The average distance between the university sites is about 3.5 km with a maximum distance of 5.1 km. The four university sites are well-connected to the public transportation network (Figure1). Three university sites can be reached without transferring between public transport modes. All transportation lines run every ten minutes, except Bus B, which runs every twenty minutes.

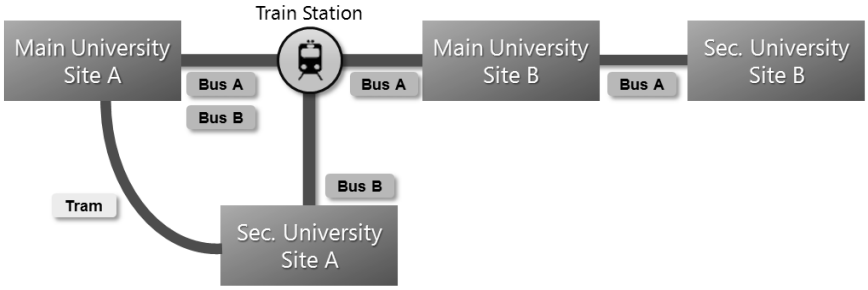


Figure 1. Location and accessibility of the sites to public transport

The need for action, i.e. to develop a multimodal mobility-on-demand system resulted out of a pre-survey, which was conducted prior to the research project. Altogether 62 employees participated (21 female and 41 male participants, in age

from 23 to 59 years). It became apparent that 44% of the employees commute between the locations more than once a day. Furthermore 64% commute at least once a week. The most commonly used vehicle for this action was a privately owned car (82%). But there are other transportation options as well. An interesting point is that 86.9% of all participants are willing to accept a longer trip time to use a carbon dioxide reduced transportation option. In detail 55.7% said that they would invest five or more minutes in transportation if there is a sustainable transportation option. On the basis of these results, relevance and potential of a multimodal mobility concept for short and mid-range distances for the employees of this university was designed and established.

In a representative survey at the start of the project, more detailed information about the current mobility behaviour was gathered in order to define the goals and the technical design of the system. The survey was split into two survey time points - summer and winter. Both surveys were conducted in term time to gather information of the mobility behaviour due to educational obligations. Aspects like the subjective assessment of mobility relevant aspects of their main location (accessibility, availability of parking slots) and overall travelling behaviour between the locations were asked. The employees were also asked their personal reasons for choosing a specific transportation option. Factors mentioned were weather, environmental friendliness, low financial effort, availability of transportation options, accessibility of parking slots and speed.

399 employees took part in the summer survey. The average age was 34 years (*SD* = 10.6), 61.7% males. In the winter survey 187 (52.8% males) employees with an average age of 34 years (*SD* = 11.0) participated. Out of a retrospective view, the participants described their daily mobility behaviour of one week. In summer, 525 trips (1.32 per employee), in winter 294 (1.65 per employee) different trips between the university sites were found. About half of the persons questioned (summer: 209 persons [52.4%]; winter: 90 persons [50.6%]) travelled once a week between the sites. The distribution of the chosen transport options, the modal share, concentrates on motorized individual transport (Table 1).

Table 1. Modal share

		Trips taken		Percentage	
		summer	winter	summer	winter
Means of transport	by foot	17	7	3.2	2.4
	Bicycle	55	13	10.5	4.4
	Private cars	319	171	60.8	58.0
	Company cars	11	2	2.1	0.7
	Public transport	112	94	21.3	31.9
	Others	11	8	2.1	2.7
	Total	525	295	100.0	100.0

The results provide a sound basis for the design of the planned mobility concept and demonstrate again the need for an alternative mobility concept. It became apparent that there is a need to change the employee's mobility behaviour from using private cars to "green" means of transportation. Through the provision of vehicles and by offering easier access to public transport through job tickets, a monetary incentive to use the mobility-on-demand system is created, since the vehicles and public transports provided are free of charge.

This monetary effort as a factor in the choice of means of transport has been queried in the survey in combination with other factors such as weather, environmental friendliness, availability of transport facilities, access to parking spaces and speed. The participants were asked to divide 100 points among the factors. Figure 2 shows, that the employees consider functional factors like availability and speed to be more important than normative aspects like sustainability.

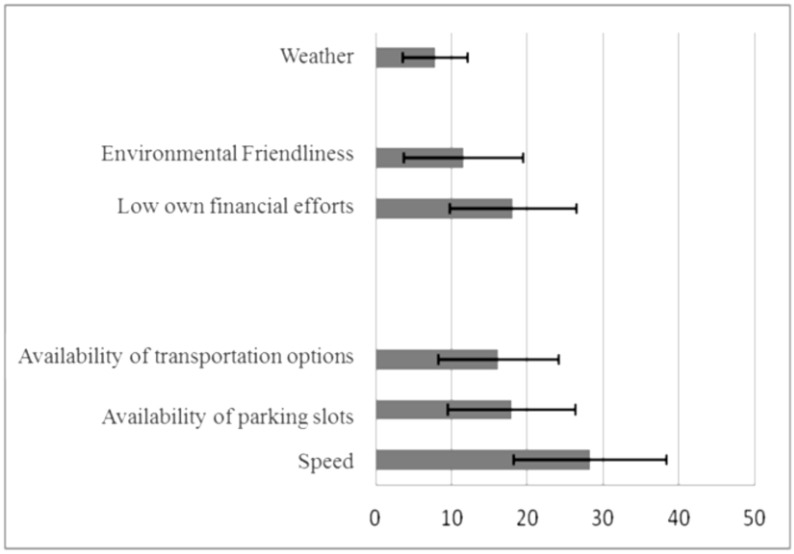


Figure 2: Characteristics in choosing transport options

Through the free provision of vehicles in the system the low-cost hypothesis, which says is that "environmental attitude affect the environmental behaviour most likely in situations that are low-cost [...] linked" (Diekmann, 1998), is fulfilled to give the user a shift from private cars to provided transport options. Matthies et al., however, criticizes, that the low-cost hypothesis does not consider the influence out of the habit (Matthies, 2006). As emerged from the survey, the use of the private cars for have a high rate (summer: 60.8%, winter: 58%) in transportation between the sites. Therefore there is a high likelihood that the user accesses only the provided electric cars. This not intended behaviour could lead to a neglect of the provided electric bikes or the public transport as alternative transport services. With this in mind, an explorative study was designed to further refine the understanding of users´ intentions and expectations towards the mobility system and its opportunities. One common definition of expectations can be found in Dorsch (2014):

Expectations are cognitions […] which express the anticipation or the forecast of future events and imply a [subjective] probability estimation of the entering of their occurrence

By this meaning, user expectations can be the cause of an actual use and also be indicative towards a frequency of use. Given this indication, the results can be integrated in the mobility system to better address users' needs and to apply solutions to systematically address unwanted user behaviour. Also keep in mind, that users´ expectation depend on the knowledge about the specific topic the expectation is about.

Data collection

From April to end of June 2014 98 people applied for participation in the field study. Due to some legal regulations only employees of the Chemnitz University of Technology were allowed to participate. Additionally it was required to agree in different ways of data acquisition during the field test. Therefore 71 applicants were selected. Their mean age was 32 years (SD=7.66) and 49 (69%) of them were male.

The selected employees were invited for participation via a telephone call. During this call a semi-structured guided telephone interview was conducted. This interview contained a questionnaire on how the participants found out about the field study (Q1: "I would like to know: how you got interested in the project?") and some questions regarding their expectations in participation. Those questions were Q2: "What do you wish or expect from participating in the project study?" Q3: "What was the main reason for participating?" and Q4: "What changes do you expect regarding your future mobility at work?" To complete the telephone interview, an appointment for an instruction regarding the handling of the mobility system and the legal conditions of its usage was made. During the interview, the audio was recorded. After the instruction participants were able to reserve and use the project vehicles. Every reservation and trip with these vehicles was recorded with reservation time, chosen vehicle, starting as well as ending point and time.

Analysis

The data analysis follows a qualitative approach after Meyring (2010) and Kuckartz (2012). The telephone interviews were completely transcribed and analysed. This study is investigating the first three questions. The answers to the questions (Q1: "knowledge", Q2: "wishes and expectation" and Q3: "main reason") were isolated and examined. To get a first impression of the text, common words were eliminated and words in direct context to the project were summarized and counted. The result, as shown in Figure 3, is most likely mobility related and targeting towards "simply testing the system/car/EV/ etc."

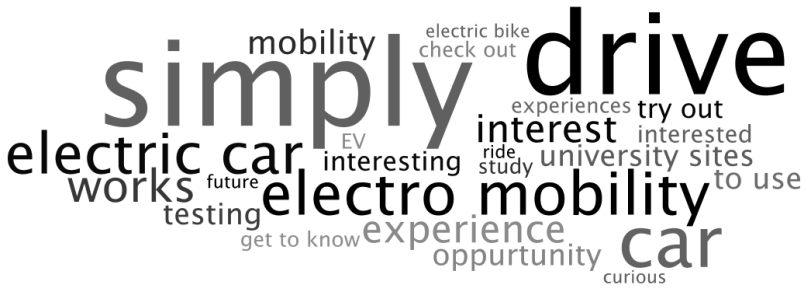


Figure 3. Characteristic wordcloud of the text

A deeper analysis with the method of the content text analysis, introduced by Meyring and developed further by Kuckartz, of the text was conducted by three raters. First step was to develop a category system to get a more plausible and easy to understandable view on the texts. The first 20 interviews were loosely categorized to get an overview over the upcoming categories. After this orienting phase, raters performed "sense-keeping" reductions to the text to understand content and meaning in a more objective way and then developed a holistic category. In two discussions, experts summarized and simplified the category system to a final stage as can be seen in table 2.

Table 2: Category system after text analysis

main category	I.	II.	III.
electromobility	a) EVs b) technology	a) electric car b) electric bike c) affinity to technology d) state-of-the-art	a) worthwhile b) driving experience c) form an opinion d) suitability for daily use e) user friendly f) driving/testing
general mobility (business trips)	c) replace own car d) leave car at home e) parking slots f) cost-effective/cheap g) fast h) comfortable i) spontaneous j) environmental friendly		
project interest	k) support project l) research interest m) interest in system	e) provide data f) be participant g) interested in results	
public interest	n) general interest o) university	h) promote general mobility i) implantation at the university	f) ecological g) economic

Four main categories with several subcategories were identified:

Electromobility

The category electromobility includes statements that include elemental propositions towards the electric vehicles (Subcategory Ia) and the technical interest (Subcategory Ib) into them. The Subcategory II differentiates into two means of electric vehicles, electric car and bike, the subjects' technology affinity and their interest in the state of the art. Subcategory III is addressing actions or intentions towards the upper categories. Typical statements are:

- *And for myself also to get a certain feeling, like how this electromobility is working.* → Electromobility, Ib, IIc, IIIf
- *[…] maybe a little more to deal with electric cars in general?* → Electromobility, Ia, IIa, IIIf
- *…so to have the opportunity at all times […]to drive a pedelec* → Electromobility, Ia, IIb, IIIb

General mobility

Main category II is summarizing statements towards general mobility issues. Subjects' expect the system to be more flexible and faster than their usual mobility solutions. They expect a cost effective and environmental friendly solution. Representative statements are:

- *…when I have a business trip, I will have an easy, unrestricted and uncomplicated solution…* → general mobility, Ig, Ih
- *…business trips […] no longer have to do with the private car…* → general mobility, Ic
- *The most important ... is, in effect, actually the parking situation* → general mobility, Ie

Project Interest

As the study is enrolled at a university, quite a large part of the statements are project related. Colleagues are interested in supporting the project as participant as well as they are interested in the results of the study. Typical statements are:

- *On the one hand to contribute to research at our university.* → project interest, Ik, IIe
- *Hmm... Actually, I'm curious how the study is structured and what I maybe can learn in my own work for user studies…* → project interest, Il, IIf
- *Yes of course they need to* [scientifically] *succeed here.* → project interest, Ik

The last main category is in general addressing normative statements. They are related to public interests to promote a change in general mobility behaviour or the subjects' concerns about the implementation of the system at the university. Statements that fit that category are listed below:

Public Interest

- *I am concerned with […] the electric mobility as a whole.*
  → public interest, Ih
- *A good appearance of the university, so economically speaking, ecologically as well.*
  → public interest, Il, IIf

For inter-rater reliability Krippendorff's alpha was calculated (Hayes & Krippendorff, 2007; Freelon, 2010). It varied between .85 and .56 (see table 3). It revealed that the first two categories "electromobility" and "general mobility" were clearly defined with a high conformity between raters. The both remaining categories "project interest" and "public interest" resulted in a middle alpha value, showing that while coding they can be interpreted broader than the first categories. Nevertheless they seem reliable enough for persisting as distinct categories.

Table 3: Interrater reliability (Krippendorff's alpha) of the main categories

main category	electromobility	general mobility (business trips)	project interest	public interest
Q2: wishes and expectations	0,85	0,72	0,66	0,62
Q3: main reason	0,72	0,69	0,61	0,56

Therefore the category system and the statements could be discussed and finally determined. Afterwards, the participants of the study were sorted into groups, which were related to mobility behaviour in system usage afterwards.

The group classification was based on the main categories built before. In Question 3 "main reason" subjects´ were able to focus one topic. Reflecting the mentioned main topic and the statements´ categorization from the answers of Question 2 "wishes and expectations", the subjects´ were divided into the following groups:

Technology driven group

The technology driven group can be described as having a general interest in electric vehicles and are also interested in the technical background of the new

technology including charging stations and typical issues like range anxiety. A high curiosity and affinity for technology as well as an interest in the state-of-the-art are characterizing for the group. 38 % of the participants could be assigned to the group.

Mobility driven group

The mobility driven group is keen on optimizing their own in-house mobility. Some of them want to replace their own car for business trips and/or want to leave their private car as a direct reaction to the implementation of the system at home. They have a high interest in a flexible and quick mobility solution which is also cost-effective and spontaneous. Out of all participants, 27 % could be assigned to the group.

Others

The Group "others" include participants who have mainly interests in the project as a scientific project. They want to provide data and/or just want to be subjects for the study. A specific main reason cannot be identified. Also environmental statements as part of a general public interest are assigned to that group. The rest of the group are participants which cannot be easily assigned to the first two groups and also do not fit into the project or public interest group.

Results

The main groups *"technology driven"* and *"mobility driven"* were now compared with gathered system data to get a distinctive view if the two groups correlate with the overall usage of the system. The data evaluated for the paper includes the booking inquiries and date of the taken business trip. The 68 participants completed overall 881 trips starting at April, 14th until October, 9th in 2014. Exceptions during data evaluation have to be made: As users did not volunteer at the same time, most of the users have individual starting points and therefore a cumulative view is not appropriate. Instead, each dataset for every user needs to be aligned relative to each other. The mean number of taken trips per month is shown in figure 4. Statistical tests showed no effect between expectation groups and long-term usage. The high variances within the expectations groups indicate that expectations are not suitable for long term usage predictions. Data analysis shows that each user group contains power users and users with less than one trip per month.

For short-term, a different result can be shown. Analysing the data on weekly basis, the expectation groups differ in their usage at early stages. The numbers shown in figure 5 indicate high usage behaviour of the "technology driven" and the "mobility driven" group. For the first weeks those groups undertake more trips per week than the group "others". After that, the two groups alter their behaviour to a more "normal" use close to 0.5 trips per week, which is the average overall system usage by all users per week.

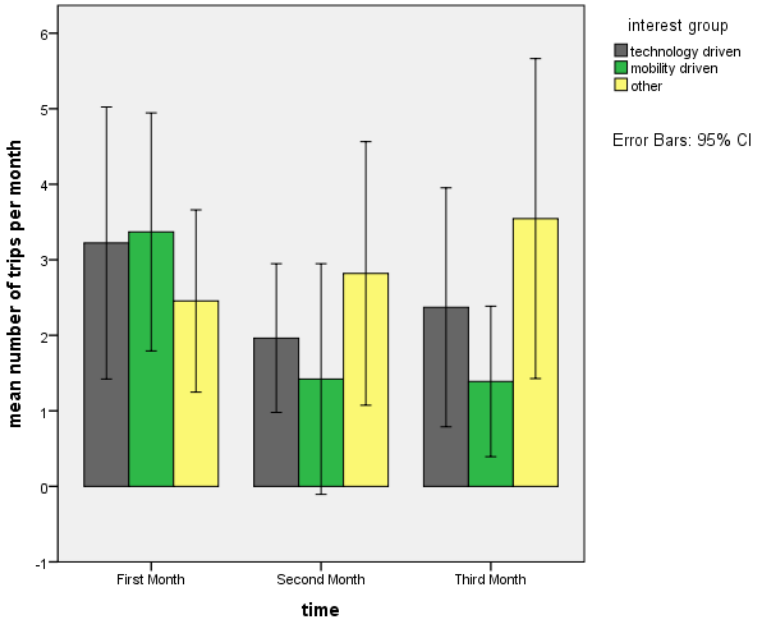


Figure 4: Mean number of taken trips per month

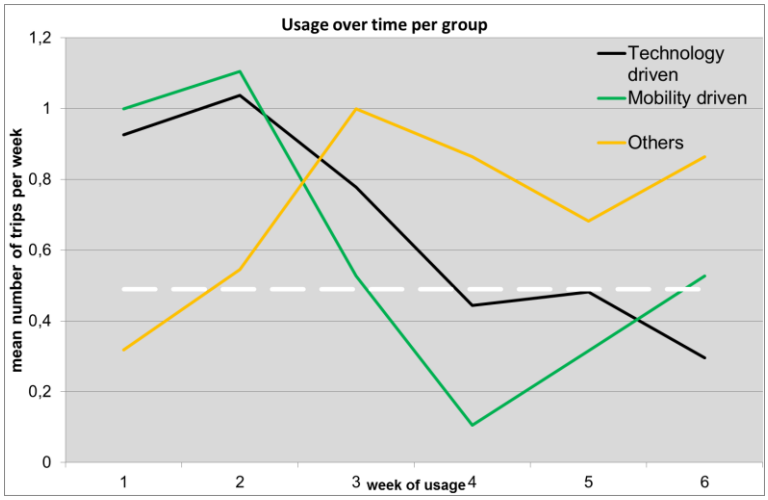


Figure 5: Mean number of taken trips per week

Conclusions

This explorative approach indicates that initial expectations are not a suitable measure for long-term usage predictions. A closer look at short-term usage seems to indicate that initial expectation of users may be linked to short term behaviour while after an initial use all groups seem to adjust their behaviour and tend to adjust also

their expectations. The ongoing research with mid-term surveys and final surveys will clarify those indications.

However, the link between users' expectations and short term behaviour shows a potential of the approach to form initial user groups. Those user groups and their expectations can be used in implementation phases of a technical system. With a deeper understanding of the relation between expectations towards technical systems, i.e. a generalized expectation model, a broader audience besides the typical early adopters can be addressed. The system can then be adjusted in a way that system will fit more users' needs than has been reached by just implementing the user-centred design process. Nevertheless, there is research potential headed for a better general understanding of expectations towards technical systems. When generalizing expectations, user groups can then be clustered into user groups for predicting their behaviour. The next steps in further research is intending to answer the question, if (initial) expectations towards a technical system are suitable for clustering user groups for predicting later users' usage behaviour.

Acknowledgements

The fahrE project is funded by the European Social Fund (ESF, 2014) of the European Union and the Free State of Saxony, Germany. The authors would like to thank ESF for supporting this work. Furthermore, the commitment of the fahrE partners Photon Meissener Technologies GmbH, EA EnergieArchitektur GmbH, MUGLER AG, GHOST Bikes GmbH, ELICON GmbH and the city of Chemnitz is gratefully acknowledged.

References

Die Bundesregierung (2009). Nationaler Entwicklungsplan Elektromobilität der Bundesregierung [National Development Plan Electro-mobility oft he federal govournment]. Retrieved: 13th Aug, 2014, from http://www.bmbf .de/pubRD/nationaler_entwicklungsplan_elektromobilitaet.pdf

Diekmann, A./Preisendörfer, P. (1998): Umweltbewußtsein und Umweltverhalten in Low- und High-Cost-Situationen. Eine empirische Überprüfung der Low-Cost-Hypothese. Zeitschrift für Soziologie 27 (pp. 438-453).

Pripfl, J., Aigner-Breuss, E., Fürdös, A. & Wiesauer, L. (2010) Verkehrsmittelwahl und Verkehrsinformation. Emotionale und kognitive Mobilitätsbarrieren und deren Beseitigung mittels multimodalen Verkehrsinformations-systemen. EKoM-Endbericht. Wien: Kuratorium für Verkehrssicherheit.

ESF: European Social Fund, Online: http://www.smwa.sachsen.de/de/Foerderung/ Strukturfonds_in_Sachsen/Europaeischer_Sozialfonds_ESF/120624.html

Kuckartz, U. (2012): Qualitative Inhaltsanalyse. Methoden. Praxis. Computerunterstützung. Beltz/Juventa. Auflage: 2., durchgesehene Aufl. (9. Januar 2014) (ISBN 978-3779929222)

Matthies, E. et al. (2006): Applying a Modified Moral Decision Making Model to Change Habitual Car Use: How Can Commitment be Effective?, *Applied Psychology: An International Review, 55,* 91-106.

Meyring, P. (2010) Qualitative Inhaltsanalyse. Grundlagen und Techniken, Weinheim. Beltz; Auflage: Neuausgabe, 11., vollständig überarbeitete Aufl. (3. September 2010) (ISBN: 978-3407255334)

Mobilität in Deutschland (2008). Ergebnisbericht Strukur – Aufkommen – Emissionen – Trends[mobility in germany, report structure – emergence – emissions – trends], 84 - 87.

Wirtz, M.A. (Hg., 2014): Dorsch - Lexikon der Psychologie. 17. überarb. Auflage. Verlag Hans Huber, Bern, 2014 (ISBN 978-3456854601)

Are globality and locality related to driver's hazard perception abilities?

Shani Avnieli-Bachar, Avinoam Borowsky, Yisrael Parmet,
Hagai Tapiro & Tal Oron-Gilad
Ben-Gurion University of the Negev, Israel

Abstract

Driving requires various skills, amongst them hazard perception that has been directly linked to involvement in traffic accidents. Navon-type tasks may provide a framework for understanding perceptual processing and logical reasoning. Yet, limited attempts were made to formulate associations between globality and locality in visual processing and perception of real world stimuli like hazards while driving. A study aimed to link Navon-type tasks with hazard perception abilities of drivers was conducted. A sample of 39 young novice drivers, 60 adult students, and 21 adult drivers completed a battery of cognitive test including Navon tasks. Then they performed a hazard perception test (HPT), in which they observed video-based traffic-scenes and were asked to press a response button each time they detected a hazard, followed by classification and rating of hazardous scenes. While there is a known statistically significant effect for experience, results reveal significant ties between global and local processing, and hazard perception. The significant effect of the global/local scores in the Navon tasks on performance on a real-world traffic situation test suggests that the Navon tasks, as well as other cognitive tests may be useful in predicting performance in real world complex situations such as driving.

Introduction

Among the different types of skills required for good driving, the only one that has been identified to have direct connection to involvement in car accidents is Hazard Perception(HP), the ability of the driver to predict hazardous situations (Horswill & McKenna, 2004). Studies have shown a connection between cognitive abilities and car accidents occurrences, mostly among adults (Horswill et al., 2008). In McKnight and McKnight (1999), cognitive abilities such as: attention allocation, perception speed and short-term memory were tested, along with their effect on driving skills and dangerous behaviour on the road. A positive correlation was found between driving skills and cognitive abilities. In addition, a negative correlation was found between the number of traffic tickets and car accidents in which the participant was involved in, and cognitive abilities. Other studies found that other cognitive skills such as: spatial perception (Maratolli, 1998), and handling functions (Daigneault, 2002) affected driving.

Several studies have shown that the ability to perceive hazards is related to Higher-Order abilities. Among them: cognitive flexibility, problem solving, urge control (Delis et al., 2001), task analysis, strategy monitoring, (Borkowsky & Burke, 1996),

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

attentional control and goal setting (Anderson et al., 2001). Horswill and McKenna (1999) have found that during a hazard perception test, young drivers who were given an additional verbal task, that was nor motor or visual, were taking substantially more time to respond to danger. They therefore claim that the hazard perception process requires allocating higher-order abilities.

In the literature the terms "Hazard Perception" and "Situational Awareness" are described as bearing one meaning when it comes to driving. They both describe the way in which a driver is aware of details in the surroundings (vehicles, pedestrians, traffic-signs), and how the awareness to those details aids him in predicting hazardous situations. Situational awareness (SA) is considered to be related to higher-order cognitive abilities. In FMRI and EEG imaging tests, a connection was found between SA and cerebral structures in charge of higher-order cognitive abilities (Borghini et al, 2012; Brookings, Wilson & Swain, 1996). Previous studies, specifically conducted on experienced drivers, detailed the cognitive and psycho-motor features that effect driving and safety (Horswill et al., 2008; Anstey, 2005), but the cognitive abilities that relate to hazard perception among all driver population is a field that has yet to be studied.

A traffic Hazard Perception Test (HPT) was developed at Ben-Gurion University of the Negev (Borowsky, Oron-Gilad, Shinar &Meir, 2010). The HPT is a computer-simulated test composed of three components: at first drivers are shown a set of movies and asked to identify hazardous situations, all while eye-motion is being documented; in the second component they are asked to classify scenes by certain similar properties, and in the final stage – a set of six pairs of still-scenes are shown and at this point they are asked to mark the still that they perceive as being the more dangerous one (see Borowsky & Oron-Gilad, 2013 for detail).

In the current study, the aim was to find ties between the HPT components and higher order cognitive abilities. Specifically two cognitive abilities were examined: logical reasoning by using Navon tasks (Navon, 1977; Stanovich & West, 1997) and attentional control, by the Attentional Control Scale (Derryberry & Reed, 2002). According to Navon (1977; Schooler, 2002, in Foster, 2010), people can use different *processing styles*. By using a global processing style, people attend to the Gestalt of a stimulus set, whereas when using a local processing style they attend to its details. The attentional control scale measures abilities such as attention focusing, directing attention from target to another and thoughts control. It was hypothesised that participants' performance in logical reasoning and attentional control will affect their scores in each one of the three components of the HPT.

Experimental materials and method

Participants

A hundred and twenty participants: 39 young-novice drivers (17–18 years old) with less than *three months* of driving experience; 60 experienced drivers (24–28), with an average of 8.2 years of driving experience; 21 very experienced drivers (40–60), with an average of 28.2 years of driving experience. Young-inexperienced participants received monetary compensation for their participation. Experienced

participants completed the experiment for course credit in an introductory ergonomics course and the very experienced drivers had all volunteered. The very experienced participants were recruited from the city of Ashkelon. The experienced drivers were students in the University, and the young-inexperienced drivers were recruited through driving schools in the city of Beer-Sheva.

Apparatus

Participants were measured for their abilities in driving related hazard perception, and other more general cognitive abilities, using a computer-based test. The hazard perception test (HPT) was the one developed by Borowsky, Meir, Oron-Gilad and Shinar (2010). Through the years several sessions of experiments were executed, and several iterations of the HPT were created in order to refine the test until it reached its final version. A 19" wide screen with 1024×768 pixels was used to display the hazard perception test and the cognitive abilities tests. Participants sat at an average distance of 70 cm from the screen.

Hazard Perception Test (HPT)

The HPT includes three components: Identification, Categorization and Rating tasks. In the identification component participants were asked to observe 21 traffic movies from the perspective of a driver and press the "Space" button on the keyboard each time they detected a hazardous situation. At the end of each movie participants were asked to verbally note the hazard instigator for each hazardous situation that they have detected (Figure 1).

Figure 16. Left: An example of the movie presentation (a snapshot sample) and Right: the following screen, at the end of the movie where participants had to register each button press they made.

Following the active identification component, in the categorization task, participants observed eight traffic scene movies for the second time and were asked to categorize them into an arbitrary number of groups according to the similarity in

their hazardous situations (Figure 2). This procedure resembled the one used in Borowsky et al. (2009).

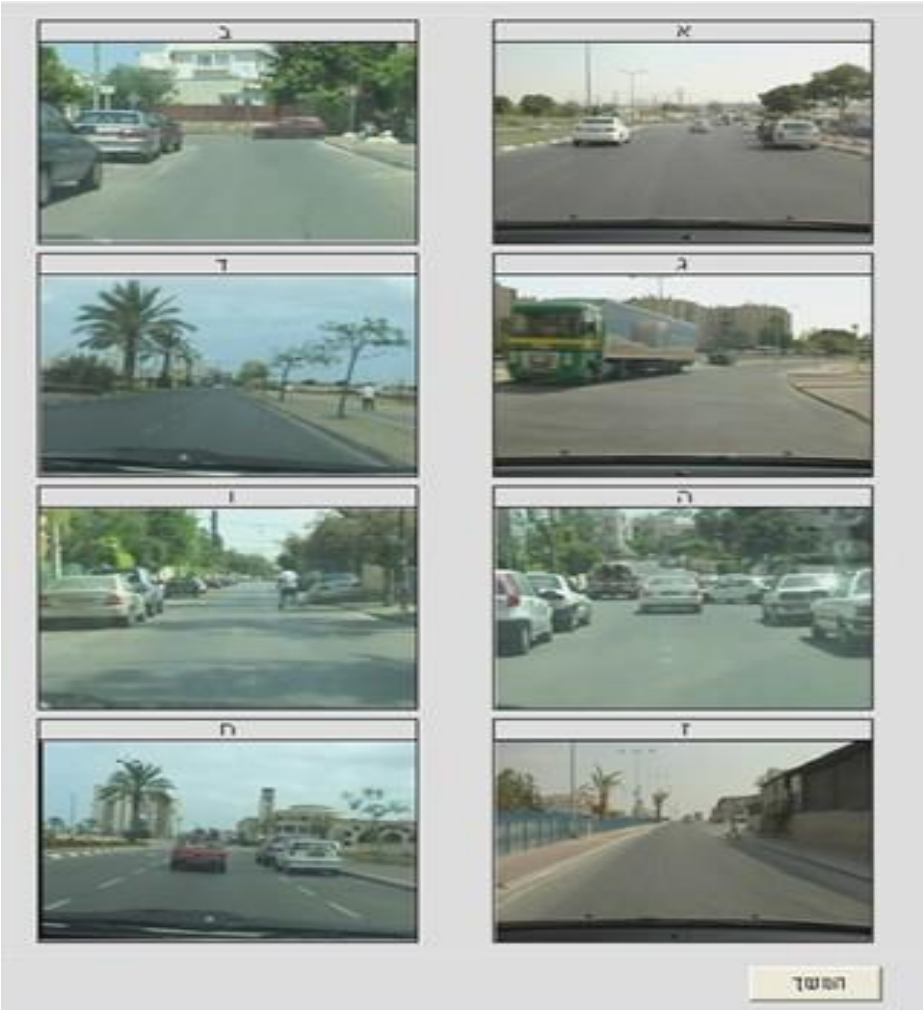


Figure 17. Representative photos of eight movies that were used in the classification component of the computer based HPT.

In this third component, participants were asked to compare 6 pairs of pictures that were taken from the HPT movie data base that the participants just observed in previous components of the test. In each comparison two pictures appeared with a hazardousness scale. The scale ranged from "a more severe danger/a greater danger" located at the end of the scale under each picture, and "equal danger" which was located in the middle between the two pictures (Figure 3).

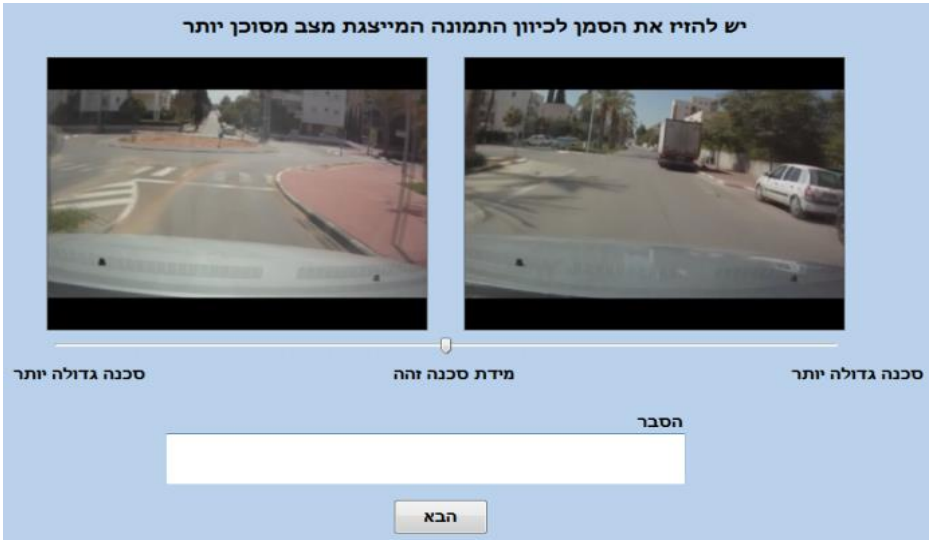


Figure 18.Sample screen of the rating task.

Cognitive abilities

The cognitive test battery consisted of Navon tasks (Navon, 1977) and the Attention control scale (Derryberry & Reed, 2002) using standard administration protocols and trained examiners.

Navon Tasks. In this type of tasks, the global large letter is combined of small letters (e.g., the letter 'H' in Figure 4).Attending to the large letter represents the global level, while attending to the small letters represent the local level. Participants were asked to press the letters 'S' or 'H' on a keyboard as soon as they identified one of those letters, when the letter could be portrayed either as a big letter (global level), or a small letter constituting a big letter (local level). In each of the conditions, the big letter is presented as slanted to one side and for a very brief time. In this task, the accuracy is measured, meaning whether the participant was correct/incorrect in identifying the letter, and the response time (RT).

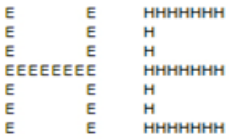


Figure 4. On the right image, the small letters 'H' represent the local level. On the left image, the letter 'H' represents the global level (Navon, 1977).

Attention Control Scale. Attention control is evaluated according to a value scale containing a person's personal ability to divert attention to the appropriate direction in accordance to the environment in which he is located. Factor analysis indicates that the scale measures the general ability to control attention, combined with the following personal skills: a) focusing the

attention, b) displacing/ refocusing attention between various tasks, c) flexibility of control in thinking.

Procedure

To generate randomization and to avoid priming effect, half of the participants started with the hazard perception test, and half of them started the experiment with the cognitive ability battery. Before starting the HPT, participants were given instructions and two training hazard perception movies in order to verify that they understood the experimental task. At the end of the training, the instruction screen appeared again, and then the test with the 21 hazard perception movies began. The order of the movies was randomised and movies were separated by a fixation screen. Upon finishing the hazard detection task, participants were instructed on the classification task. They were told that they are about to observe eight movies for the second time, and then were asked to name each group of movies, in a way thought would best describe them. In the third part of the test participants were asked to compare 6 pairs of still-scenes and to locate the pointer on the hazardousness scale.

Before starting the cognitive tests, participants were told that they would be completing a series of cognitive ability tests. The presentation order of the two was counterbalanced across testing sessions. Written test instructions were read aloud to the participants by the test examiner before each test was administered. Participants completed the cognitive ability battery and the three components of the HPT, in a total duration of approximately 1 hour.

Results

The main purpose of the experiment was to discover whether having specific abilities and traits related to visual attention, can predict perception of real world stimuli like hazards while driving. The analysis was made on the three components of the HPT. Results are presented in the same order as the experimental procedure. The generalised linear mixed-effects model (GLMMs) was used for the statistical analysis. Performance outcomes of the three hazard perception's components were set to be the explanatory variable; Navon tasks, and the attention control scale were set to be the potential exploratory variables of the model. The 'participants' variable was set to be the random effect, and included in the model in order to care for individual differences among participants. Using a backward elimination process by p-values, the most fitting model was set for each of the explanatory variables. Consequently, the appropriate GLMM was applied on the data set.

Component 1identification of events in a dynamic scene

Accuracy. Hazardous events were not defined a priori, but were data driven, subjectively defined according to the pool of all participants' responses. The beginning of an event was defined as the minimum time to respond, of all responses related to it. Similarly, the end of each event was defined as the maximum time to respond, of all responses related to it. Thus, the duration of each event was defined as the time interval between its beginning and its end. Participants' responses which referred to the same hazard instigator and had a temporal proximity to each other

were gathered and defined as the same event. I.e., an event was an instance that was detected by any number of participants, who used approximately the same explanation to describe it, and occurred in nearby frames. Each of these events was then titled based on the idiosyncratic definitions given to it by each of the participants who registered it. This event-definition procedure ended up with a total of 67 events spread across the 21 HP test movies. Events were defined and labeled as "*Experienced-Based Events*"(EBEs) if at least 31 of the experienced drivers group (i.e., 50% of the experienced drivers) reported them as hazardous (as reflected by their button presses and written descriptions). This criterion allowed the creation of an array of representative, noteworthy genuine EBEs, thus enabling the experienced-drivers' group to be set as a goal standard.

A multinomial regression with a logit link function was applied in the framework of GLMM. The dependent variable was response to EBE (0 or 1) and the independent variables were (1) Logical reasoning measured by Navon task measures and 2) Attention control by attentional control (AC) scale. Applying a backward elimination procedure found the best fitting model has two significant exploratory effects: Accuracy and Global RT in the Navon tasks both were statistically significant ($F(1,111)=10.39$, $p=.002$; $F(1,111)=7.17$, $p=.008$), respectively. No significant effect has been revealed for AC. Meaning, there was a correlation between recognition of hazardous events, and logical reasoning, and the processing of global features as they show in the Navon tasks.

Response time. Response time analysis was also conducted on the EBEs, where there is meaning to the immediacy of response. For each of the participants' mouse clicks, an elapsed time in milliseconds was assigned for the time in which it was performed since the initiation of the video. Each press made by a participant was recorded according to its frame number. To standardize response time, the interval between a participant's response (frame number) to a specific event and the beginning of that event (frame number) was divided by the length of the event (frames), see Equation 1.

Applying the GLMM revealed that AC, and Navon in the Global RT were statistically significant predictors for quicker reaction time to hazardous events ($F(1,113)=29.38$, $p<.0001$; $F(1,113)=19.57$, $p<.0001$; $F(1,113)=10.97$, $p<.0001$, respectively). Meaning that there was a correlation between fast responses to hazardous events, and logical reasoning and global analysis and AC.

$$Standardized\_response\_time = \frac{Participant's\_response\_frame - event\_start\_frame}{Event\_end\_frame - Events\_start\_frame} * 100$$

Equation 1. Calculation of the standardised response time for EBE events in the computer based HPT component 1

Component 2 – Classification

In the categorization task the dependent variable, i.e., the number of categories and the number of movies in each category varied from one participant to another. Although the number of possibilities to categorize the movies is theoretically

unlimited, the observed number of arrangements was much smaller (e.g., Borowsky et al., 2009). It was decided to identify dominant clusters of movies), i.e., clusters or combinations of clusters that were categorised by a certain percentage of all participants and to scale them according to the abstraction of the classification, where higher abstraction implies on higher understanding of the road environment. Five possible structures for classification were rated with 1 being a criterion of classification indicating a low level of classification abstraction and 5 being the highest level according to Benda and Hoyos (1983): (1) No hazards; (2) Similarity in the Hazard Instigator - intersection, pedestrian, field of view, other vehicle behavior, driver's behavior, crosswalk and traffic circle; (3) Hybrid-Hazard Instigator and Traffic Environment ; (4) Traffic Environment- urban, residential, and inter-city; (5) Level of hazard-low, medium and high, A priori, the set was categorised on the basis of these five structures. This a priori categorization reflects exclusive reliance on either one of the five structures– that is a driver who related solely to a single categorization criterion (1-5). Notably, it was not expected that drivers will categorize movies exclusively according to the pre-defined categorization structures (Ahn & Medin, 1992).

Multinomial regression with a Logit link function was applied in the framework of the GLMMs on the data. Applying a backward elimination procedure found the best fitting model had only one statistically significant exploratory effect: Navon task in the local Accuracy condition ($F(1,111)=3.039$, $p=.08$). Meaning, there was a correlation between participants who succeeded in the classification task who their abstraction rank was high and logical reasoning and local analysis.

Component 3 – Ratings

In this task each participant was asked to compare between two pictures by locating the pointer on the picture according to its danger. Analysis of the results began by defining a priori a gold standard solution based on a pilot experiment with very experienced drivers. Based on its results every one of the six comparisons got one of three grades: The highest grade was 2, meaning that the selected picture was rated as more dangerous than the other picture according to the gold standard solution; The lowest grade was 0, meaning that the selected picture wasn't the most dangerous picture out of the two pictures according to the gold standard solution; The intermediate grade 1-was given when these two pictures were equally dangerous. After grading each of the six comparisons for every participant, an average grade was calculated for each participant's rating abilities. The minimum average grade was 0.33 and the maximum average grade was 2, which means that some participants were always correct in the way they rated the most dangerous picture in a similar way to the gold standard solution. Applying the GLMM revealed that only the Navon task in the Local RT condition affects the correct rating in comparing two hazardous events ($F(1,113)=10.45$, $p=.002$). Meaning, there was a correlation between successful rating of hazardous events, and logical reasoning and local analysis.

Discussion

In the literature there is little evidence of studies that tested which of the higher-order cognitive abilities may improve drivers' ability to perceive hazards. The current study is a primary research to evaluate the connection between logical reasoning, and attentional control, and the ability to commit hazard perception in driving.

Logical reasoning was measured using the Navon Tasks. In the Navon tasks several sub-measures were produced: an accuracy scale – whether the right letter was recognised, a response-time scale – how fast was the letter recognised, a scale for the letter's level – whether the recognised letter was in a global or a local processing level. It was found that logical reasoning was manifested in all three components of the HPT; participants who succeed in the identification task, and detected many events quickly, succeed as well in logical reasoning with global analysis. Furthermore, participants who succeed in the classification task, and their abstraction rank was high, succeed as well in logical reasoning with local analysis. In addition, Participants who succeed in the rating task, and their rating's score was high, succeed as well in logical reasoning with local analysis.

Navon (1977) has claimed that people process information using two procedures: a global processing procedure and a local processing procedure. At the global level, processing is conducted by a general stimulations layout (looking at the "forest"), while at the local level, it is conducted by a more specific layout paying attention to details (looking at the "trees"). Navon has demonstrated that the time it takes to respond to a big letter (global level) is shorter than the time it takes to respond to the smaller letters (local level) that make the bigger one. With this in mind – he claimed that the entire population's default is the global processing procedure. Based on Forster's (2010) research and Navon's (1977) claim that the population's default is indeed a global processing procedure – it could have expected that the processing procedure in the Navon Task would be global. The results of the study have shown a different outcome: while at the first component of the hazard perception test – the corresponding processing procedure in the Navon Task was global, as expected, for the second and third components - classifying and rating pictures - the processing procedure in the Navon Task was local. This can be explained by the characteristics of the HPT. For the first component of the hazard perception test – each movie scene evolved and changed in a short period of time – at which the participant was required to identify the danger and respond by pressing a button. In the second and third components – the time to respond was unlimited. When participants are required to perform the test component in an unlimited amount of time –there is no feeling of pressure– as opposed to the first stage. Therefore, the test participants performed the classifying and rating tasks while delving into details and concentrating on all of the pictures' elements

As for the attentional control, it was found that this ability was reflected only in the first component of the test and not in the second and third components. There is disagreement in the literature as to how this ability is expressed in dynamic versus static displays. Some researchers claim that attentional control reflects differently in

dynamic displays as opposed to static ones (Kramer, Larish & Strayer, 1995). It is possible that the characteristics of the first component of the test, movies that dynamically changing, compared to the second and third components which are presented in a static display, contributed to the way this cognitive abilities were expressed in the test.

Research limitations and practical implications

According to Horswill (2008), a decline in cognitive abilities can affect the manner in which adult drivers perceive hazards. Studies show a decrease in responsiveness (Salthouse, 1996) and task-shifting (Mayr & Liebscher, 2001) that can affect a driver's performance on the road. This calls for a future research that will focus on adults above the age of 65, and test how hazard perception takes place among that particular group, considering sustained changes in cognitive procedures.

In the current research - a computed hazard perception test was used. It consisted of 21 videos depicting various hazardous situations. This study cannot reflect upon the total spectrum of driving situations, nor can it simulate all realistic driving situations. Additionally – higher-order cognitive skills include: problem solving, rule activation, attention, locating and fixing errors and memory. Perhaps other cognitive abilities would render different results than the current ones. Since this study constitutes a primal research, future studies can measure the effects of the aforementioned higher-order cognitive skills on drivers' ability to perceive hazards.

Neuroergonomics is a field that has evolved during the past several years and it holds two principles: Neuroscience and Ergonomics. One of the purposes of Neuroergonomics is to establish and expand an understanding of the connection between brain functions and real-life performances (Parasuraman & Rizzo 2007; Parasuraman, 2003; 2008). Due to the emergence of un-intrusive brain-monitoring techniques, future studies can test the functionality of specific areas that act during hazard perception in driving, and by doing so – determine the location of that specific area of the brain and the set of cognitive skills involved.

Mapping and identifying the cognitive abilities required to perceive hazards may be beneficial at two levels. First of all – from the evaluation side –a hazard perception test can be used to assess the performance of a driver on the road, specifically among the senior adult community – where there is a need for more assessment tools to determine competency. A computerised HPT that also measures cognitive abilities such as spatial ability, logical reasoning and attentional control can objectively assess a person's competence to drive a vehicle, as opposed to today's subjective evaluation. Secondly – since it was found that cognitive abilities have an effect on hazard perception in driving, a training program devised to improve theses abilities can be issued, so that weaker populations such as senior citizens and people with attentional disabilities can train to improve their ability to perceive hazards.

References

Ahn, W.K., &Medin, D.L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*, 81-121.

Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child neuropsychology*, *8*(2), 71-82.

Anstey, K. J., Wood, J., Lord, S., & Walker, J. G. (2005). Cognitive, sensory and physicalfactors enabling driving safety in older adults. *Clinical Psychology Review, 25*, 45–65.

Benda, H.V., & Hoyos, C.G. (1983). Estimating hazards in traffic situations, *Accident Analysis & Prevention, 15*, 1-9.

Daigneault, G., Joly, P., & Frigon, J.Y. (2002). Executive functions in the evaluation of accident risk of older drivers. *Journal of Clinical and Experimental Neuropsychology*, *24*, 221-238.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2012). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews, 44*, 58-75

Borkowski, J.G. & Burke J. (1996).Trends in the development of theories, models, and measurement of executive functioning: Views from an information processing perspective In G.R. Lyon and N.A. Krasnegor (Eds.), *Attention, memory, and executive functioning* (pp. 235–262). Baltimore: PH. Brookes

Brookings, J.B., Wilson, G.F., & Swain, C.R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology, 42* 361-377.

Borowsky, A., Oron-Gilad, T. & Parmet, Y. (2009). Age and Skill differences in classifying hazardous traffic scenes, *Transportation Research part F, 12*, 277-287.

Borowsky, A., Meir, A., Oron-Gilad, T., Shinar, D., & Parmet, Y. (2010). The effect of hazard perception training on traffic-scene movies categorization. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, No. 24, pp. 2101-2105). Sage Publications.

Borowsky A., & Oron-Gilad T. (2013). Exploring the Effects of Driving Experience on Hazard Awareness and Risk Perception via Real-Time Hazard Identification, Hazard Classification, and Rating Tasks. *Accident Analysis and Prevention, 59*, 548–565.

Delis, D., Kaplan, E., & Kramer, J. (2001). *Delis–Kaplan executive function system technical manual*.San Antonio, TX: Psychological Corporation.

Derryberry, D., & Reed, M.A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of abnormal psychology*, *111*, 225-236.

Förster, J., & Dannenberg, L. (2010). GLOMOsys: A systems account of global versus local processing. *Psychological Inquiry*, *21*, 175-197.

Horswill, M.S., & McKenna, F.P. (1999). The Effect of Perceived Control on Risk Taking1. *Journal of Applied Social Psychology*, *29*, 377-391

Horswill, M.S., & McKenna, F.P. (2004). Drivers' hazard perception ability: Situation awareness on the road. In S. Banbury. & S. Tremblay (Eds.), A cognitive approach to situation awareness (pp.155-175). Aldershot: Ashgate

Horswill, M.S., Marrington, S.A., McCullough, C.M., Wood, J., Pachana, N.A., McWilliam, J., & Raikos, M.K. (2008). The hazard perception ability of older drivers. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *63*, P212-P218

Kramer, A.F., Larish, J.F., & Strayer, D.L. (1995). Training for attentional control in dual task settings: a comparison of young and old adults. *Journal of Experimental Psychology: Applied*, *1*, 50-76.

Marottoli, R.A., Richardson, E.D., Stowe, M.H., Miller, E.G., Brass, L.M., Cooney Jr, L.M., &Tinetti, M.E. (1998). Development of a test battery to identify older drivers at risk for self-reported adverse driving events. *Journal of the American Geriatrics Society*, *46*, 562-568.

McKnight, A.J., & McKnight, A.S. (1999). Multivariate analysis of age-related driver ability and performance deficits. *Accident Analysis & Prevention*, *31*, 445-454.

Mayr, U., &Liebscher, T. (2001). Is there an age deficit in the selection of mental sets?. *European Journal of Cognitive Psychology*, *13*, 47-69

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, *9*, 353-383.

Parasuraman, R. (2003). Neuroergonomics: research and practice. *Theoretical Issues in Ergonomics Science*, *4*, 5-20.

Parasuraman, R., & Rizzo, M. (2007). *Neuroergonomics: The brain at work*. Oxford ; New York: Oxford University Press.

Parasuraman, R. (2008). Putting the brain to work: Neuroergonomics past, present, and future. *Human Factors*, *50*, 468-474

Reynolds, J.H., &Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611-647.

Salthouse, T.A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological review*, *103*, 403-428.

Stanovich, K.E., & West, R.F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, *89*, 342-357.

How to improve training programs for the management of complex and unforeseen situations?

Marie-Pierre Fornette, Françoise Darses, & Marthe Bourgy
Institut de Recherche Biomédicale des Armées
France

Abstract

Maintaining adequate performance in the face of complex and unforeseen situations is of fundamental importance in aeronautics. Such situations are often ill-defined. Therefore pilots must first determine which aspects of the situation are relevant to process and control. One major difficulty stems from the fact that this process of "situation structuration" must be performed on the basis of current constraints rather than preconceived knowledge. Thus, the key question is "What to process and control?". Currently, most unforeseen-situation management training programs do not help pilots to answer this question. Rather, by improving the ability to control the relevance of thought processes, they concentrate on another question "How to process and control?". Recent studies on thinking dispositions (Stanovich, 2011) and on mindfulness (Kabat-Zinn, 2003) are opening new avenues for training. By focusing on the development of openness and acceptance attitudes, these approaches could help pilots to efficiently structure complex and unforeseen situations. We present studies carried out in risky work environments, the results of which indicate that trainings that seek to foster an open state of mind provide a necessary complement to trainings centered on the control of thought processes, to improve pilots' ability to manage complex and unforeseen situations.

The management of the complex and the unforeseen among pilots

Aeronautic environments are traditionally characterized by multiple (physiological, psychological, and organizational) constraints. Moreover, during the last decade, the context in which military pilots operate has changed dramatically, becoming at the same time more complex and more unpredictable. Technological innovations, restructurings, and the ever-increasing complexity and diversity of airborne systems and military operations require from pilots that they be able to deal with highly complex and often unforeseen situations.

In this context, it is important to examine how pilots are able to handle such situations. In a recent study, Casner, Geven, and Williams (2013) confronted airline pilots with three abnormal events: (a) aerodynamic stall, (b) low-level wind shear, and (c) engine failure on takeoff. Each of these events was presented to pilots in two different ways under: (a) the familiar circumstances used during airline training, or (b) unexpected circumstances, as might occur during a flight. The results showed that, for approximately one third of the pilots, performance was severely hampered

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

when the event occurred in unusual circumstances. In the context of military aviation, a recent study by Bourgy (2012), in which fighter pilots faced an unforeseen situation in a simulator, found a similar proportion of failures: one third of the pilots failed at grasping the dysfunctions that they encountered, leading them to eject in a rushed and dangerous manner. Only two thirds of the pilots avoided such an unsatisfactory ending owing to their use of adaptive solutions. Moreover, an analysis of recent reports from the French Defense Air Accident Investigation Board reveals that some pilots are unable to use adequate adaptation skills to deal successfully with complex and unforeseen situations (BEAD-Air, 2004, 2006, 2007). In particular, pilots failed to recognize and understand the high stakes involved in these situations, or to take into account all of the constraints associated with the situation when making decisions.

Training pilots to better deal with unforeseen circumstances is increasingly being recognized as a need by the aeronautics community. In 2002, the French Air and Space Academy acknowledged that the training of civil-aviation pilots was incomplete because it did not sufficiently train pilots to cope with unforeseen situations (AAE, 2002). In 2011, the same academy organized a colloquium entitled "Air transport pilots facing the unexpected" (AAE, 2013), the aim of which was to survey ways of improving the management of complex and unexpected situations at the organizational, team, and individual levels. During this colloquium, colonel Rabeau (2011) pointed out that "the missions of [military] pilots in hostile environments, by nature, involve the unexpected" (p. 115), and that the training curriculum of military pilots seeks to prepare them for this by taking into account, not just technical skills, but also "an ability to step back from the mission, as well as analytic and decisional abilities" (p. 119). However, we suggest that, to further improve the training of pilots, it is essential to try to better understand the processes underlying the ability to manage unforeseen situations.

How to train pilots to manage complex and unforeseen situations?

What is at stake in complex and unforeseen situations?

In the studies cited above (Bourgy, 2012; Casner et al., 2013), the observed differences in performance could not be explained by differences in expertise because the pilots who participated in these studies all had the same high level of qualification. Complex and unforeseen situations that call for prompt responses seem to fall outside the scope of pilot's immediate expertise. These situations cannot be processed solely on the basis of fast associations and of easily applicable procedures. Achieving cognitive adaptation involves gathering situational cues, noticing patterns, activating relevant knowledge and heuristics, adapting strategies, and learning from the results of action (Ployhart & Bliese, 2006; Schunn & Reder, 1998). Pilots must be able to recognize atypical situations, for which no easily applicable procedure exists. They must be able to accept the unknown, and the fact that there does not always exist a predefined pattern which they can rely upon. In addition, in order to "structure" the situation, pilots must be able to grasp relevant aspects of it, even when those aspects are not salient. One major difficulty stems from the fact that this "structuration" process must take into account current constraints associated with the situation; it cannot be performed solely on the basis

of pre-established knowledge. In the face of complex and unforeseen situations, one of the key issues seems to be "What to process & control?"

Traditional training programs focused on "how to process and control?"

Today, most unforeseen-situation management training programs focus on the question "How to process and control?", which is also essential to cope efficiently with complex and unforeseen situations. These trainings usually concentrate on enhancing cognitive control. They aim at improving the conscious and deliberate regulation processes used by individuals to check the validity of their representations and cognitive processes, in order to improve decision-making or stress management.

For example, Helsdingen, van den Bosch, van Gog, and van Merriënboer (2010) proposed a training based on *Critical Thinking Instruction*. This training provides operators with a formalized questioning scheme for looking at the relevance of their cognitive processes and representations used to manage a situation. Another type of training aims to familiarize operators with reflexivity in order to lead them to think critically upon their practices (see for instance Decision-Making Training, Chauvin, Clostermann, & Hoc, 2009). Moreover, various stress-management techniques (Driskell, Salas, Johnston, & Wollert, 2008) can be taught to pilots to help them to efficiently manage the stress experienced in complex and unforeseen situations. Some techniques are based on physiological control, such as relaxation and biofeedback, to gain control over negative stress reactions (Orasanu & Backer, 1996). Other techniques are based on cognitive control or cognitive change (Gross, 2002). They seek to improve access to more adaptive thinking modes or representations by teaching pilots metacognitive techniques, such as cognitive restructuring. These techniques, which have been first introduced in clinical psychology, have demonstrated their efficacy in occupational settings (for review, Richardson & Rothstein, 2008) as well as in military personnel (e.g., Cohn & Pakenham, 2008).

By improving the ability to control the relevance of thought processes, these training programs help operators to answer the question "*How* to process and control?". However, in complex, real, and unexpected situations, it is difficult to determine rapidly and precisely, based on prior knowledge or cues, what to focus attention and control capacities onto, in other words, to answer the question "*What* to process and control?".

New approaches to improve the management of complex and unforeseen situations

Recent studies concerned with "thinking dispositions" (Stanovich, 2011) and with "mindfulness" (Kabat-Zinn, 2003) are opening new avenues for training. By focusing on the development of openness and of acceptance attitudes, these approaches could help pilots to efficiently structure complex and unforeseen situations. They may also reinforce abilities that were identified at the "Air transport pilots facing the unexpected" (AAE, 2013) colloquium. Experts who participated in this meeting suggested that pilots should be trained to learn to: (1) accept to be surprised and to face unknown and uncertain circumstances, (2) be open to new experiences, (3) know how to act outside of predefined procedures, and (4) beyond

trainings focused on task or situation management, develop trainings focused on general adaptation skills (which are useful for all tasks and situations).

New approaches and training designs for the management of complex and unforeseen situations

The concept of thinking dispositions

In 2011, Stanovich proposed to distinguish two aspects of adaptation to complex and unforeseen situations. On the one hand, there are the executive processes which allow effective processing of information identified as relevant, referred to as the *algorithmic mind*. On the second hand, reflective processes allow the individual to structure a situation, to assign meaning to it, and to build relevant frameworks given, not only the particulars of the situation, but also, the individual's own goals, values, and priorities (referred to as the *reflective mind*). The former processes address the "*How* to process and control" question, whereas the latter address the "*What* to process and control" question. According to Stanovich, processes underlying the reflective mind depend on individual characteristics referred to as *thinking dispositions*. The notion of thinking dispositions denotes a state of mind, tightly related to different individual cognitive propensities, such as: dogmatism and absolutism, actively open-minded thinking and openness, need for cognition, flexible thinking, or belief identification (Stanovich, 2011). Thinking dispositions refer to the way in which an individual interacts with the world. They predict inter-individual differences in complex reasoning tasks (e.g., Stanovich & West, 2008). These dispositions might play a crucial role in helping pilots formulate relevant goals and thinking frameworks in complex and new situations.

Training programs integrating thinking dispositions

To our knowledge, few studies have examined the effects of training programs that seek to promote thinking dispositions that favoring adaptation to complex and unforeseen situations. A first study of the effects of this type of training on flight performance and stress management was carried out in French Air Force pilot cadets (Fornette at al., 2012). The proposed cognitive-adaptation training is called *Mental Mode Management* training (Fradin, Aalberse, Gaspar, Lefrançois, & Le Moullec, 2008; Fradin, Lefrançois, & El Massioui, 2006). It aims at improving adaptation capabilities in occupational settings by allowing participants to question, and possibly, to modify their relationship with, complex and stressful situations. This training had two goals: firstly, to increase participants' awareness of their "mental mode" (i.e., the state of mind with which they approach a situation); secondly, to develop thinking dispositions such as open-mindedness, and attitudes of acceptance, nuanciation, relativization, rationality, and individualization. The results of the study suggest that this training has beneficial effects (a) on flight performance of cadets who had more difficulties during flights than other cadets, and (b) on stress management of all cadets who attended the training (Fornette et al., 2012).

The concept of mindfulness

Mindfulness refers to a state of consciousness in which an individual directs their whole attention on their present experience, internal and external, with an accepting state of mind, i.e., avoiding as much as possible reacting to the experience or judging its contents (Brown & Ryan, 2003). Mindfulness is a state of mind that can be developed with training. It seems particularly relevant in complex and unforeseen situations. Indeed, in such situations where multiple sources of uncertainty exist, including uncertainties concerning the relevant analysis framework, an open state of mind is undeniably advantageous (Dane, 2011). In the context of high-reliability organizations, Weick and Sutcliffe (2006) estimate that mindfulness is useful for managing unexpected situations because it encourages individuals to (a) keep in touch with deviating elements, (b) not distort reality to make it conform to available concepts, and (c) identify automatic reactions and associations.

Mindfulness training programs

Used initially for stress reduction or chronic pain management in patients, mindfulness training programs have progressively been adopted by healthy individuals (Grossman, Niemann, Schmidt, & Walach, 2004). Numerous studies indicate beneficial effects on cognitive functions (e.g., attentional capacities, cognitive flexibility) and emotions (e.g., mental health, emotional balance). Mindfulness trainings are now offered in professional environments. They have been shown to have beneficial impacts on several areas of work performance, such as learning, safety culture, conflict resolution, creativity, and decision-making (Passmore, 2009).

More recent studies have introduced and evaluated mindfulness training in risky environments, particularly, in military environments. For example, Jha, Stanley, Kiyonaga, Wong, and Gelfand (2010) proposed a new training program (the Mindfulness-based Mind Fitness Training) for improving operational effectiveness and building resilience to stressors in a high-stress military pre-deployment context. The evaluation of this training showed beneficial effects: (a) increases in working-memory capacity and positive affect, and (b) decreases in negative affect and perceived stress (Jha et al., 2010; Stanley, Schaldach, Kiyonaga, & Jha, 2011). However, these positive impacts were observed only for military participants with high mindfulness-training practice time. In Norway, Anders Meland has carried out studies to test the effects, and the transferability, of mindfulness training in military pilots. Preliminary results of a first study in a military F-16 fighter squadron indicate that 12-month mindfulness training is sufficient to further develop concentration and arousal regulation in individuals who already score high on such skills (Meland, Fonne, & Pensgaard, 2012). This training may also be used to protect against future functional and relational impairments that are often associated with high-stress contexts. However, it can have negative effects for participants who lack sufficient motivation to perform the training. On the basis of these preliminary results, a shorter (3-month), more targeted mindfulness training was developed and new studies evaluating its effects on cognitive function and stress among another sample of military pilots are ongoing.

Conclusion

Considering the ever-increasing constraints, demands, and changes that pilots are faced with, it seems especially important to help them improve their ability to cope with complex and unforeseen situations. New training approaches based on mindfulness or on thinking dispositions emphasize the importance for operators to be "present" in the situation, while at the same time developing attitudes of openness and acceptance toward unfolding events. In this way, an ability to see what is really present, independently of, or beyond, what is made salient by pilots' expertise can be acquired. Bottom-up processes used by operators allow them to answer the question: "What aspects of the situation should I be processing and controlling?". As a result, they become more likely to effectively "structure" and manage unforeseen and complex situations. By contrast, traditional training programs focus on reinforcing top-down processes by providing cognitive schemas that help pilots to control the thought processes that they use to manage situations. These traditional training programs seek primarily to reinforce cognitive control, whereas the new training approaches promote a "let-go" attitude. These new approaches belong to the category of "general skill training" programs. Indeed, once acquired, attitudes of openness and acceptance can be applied in all situations. The experts who met at the "Air transport pilots facing the unexpected" colloquium mentioned above recommended that the development of "general adaptation skill" trainings must be made a priority over the development of trainings focused on task management.

Even though the new training approaches described above seem very promising, to date, few studies have investigated the effects of such trainings on the management of unforeseen and complex situations in risky environments such as aeronautics. Additional studies are needed to better understand how these trainings operate, and also, how they can be better adapted to pilots. Providing pilots with trainings that are adapted to the specificities of their profession is a key step toward motivating them to practice the techniques that are taught to them in such training programs; without practice, such trainings cannot be effective (Jha et al., 2010). If future studies confirm the preliminary results obtained thus far, new training approaches that seek to foster an open state of mind could be an efficient and necessary complement to trainings centered on the control of thought processes, to improve the ability of pilots to manage complex and unforeseen situations.

References

AAE. (2002). *La formation des pilotes [Pilot training]* (Dossier No. 20). Toulouse, France: Académie de l'air et de l'espace.

AAE. (2013). *Le traitement de situations imprévues en vol [Dealing with unforeseen situations in flight]* (Dossier No. 37). Toulouse, France: Académie de l'air et de l'espace.

BEAD-Air. (2004). *Rapport public d'enquête technique : BEAD-A-2004-001-A.* Brétigny, France: Bureau Enquêtes Accidents Défense Air.

BEAD-Air. (2006). *Rapport public d'enquête technique : BEAD-air-A-2006-12-A.* Brétigny, France: Bureau Accidents Défense Air.

BEAD-Air. (2007). *Rapport public d'enquête technique : BEAD-air-A-2007-008-I.* Brétigny, France: Bureau Accidents Défense Air.

Bourgy, M. (2012). *L'adaptation cognitive et l'improvisation dans les environnements dynamiques [Cognitive adaptation and improvisation in dynamic environments]*. Thèse de doctorat en psychologie cognitive, Université de Paris 8, Saint-Denis, France.

Brown, K.W., & Ryan, R.M. (2003). The benefits of being present: mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*, 822-848. doi: 10.1037/0022-3514.84.4.822

Casner, S.M., Geven, R.W., & Williams, K.T. (2013). The effectiveness of airline pilot training for abnormal events. *Human Factors, 55*, 477-485. doi:10.1177/0018720812466893

Chauvin, C., Clostermann, J.-P., & Hoc, J.-M. (2009). Impact of training programs on decision-making and situation awareness of trainee watch officers. *Safety Science, 47*, 1222-1231. doi:10.1016/j.ssci.2009.03.008

Cohn, A., & Pakenham, K. (2008). Efficacy of a cognitive-behavioral program to improve psychological adjustment among soldiers in recruit training. *Military Medicine, 173*, 1151-1157.

Dane, E. (2011). Paying attention to mindfulness and its effects on task performance in the workplace. *Journal of Management, 37*, 997-1018. doi: 10.1177/0149206310367948

Driskell, J.E., Salas, E., Johnston, J.H., & Wollert, T.N. (2008). Stress exposure training: An eventbased approach. In P.A. Hancock, and J.L. Szalma (Eds.), *Performance under stress* (pp. 271–286). Aldershot, UK: Ashgate.

Fornette, M.-P., Bardel, M.-H., Lefrançois, C., Fradin, J., El Massioui, F., & Amalberti, R. (2012). Cognitive-adaptation training for improving performance and stress management of airforce pilots. *The International Journal of Aviation Psychology, 22*, 203-223. doi:10.1080/10508414.2012.689208

Fradin, J., Aalberse, M., Gaspar, L., Lefrançois, C., & Le Moullec, F. (2008). *L'intelligence du stress [Intelligence of stress]*. Paris, France: Eyrolles.

Fradin, J., Lefrançois, C., & El Massioui, F. (2006). Des Neurosciences à la Gestion du Stress devant l'Assiette. [Eating and managing stress with the help of neurocognitive therapy]. *Médecine et Nutrition, 42*, 75-81.

Gross, J.J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology, 39*, 281-291. doi:10.1017.S0048577201393198

Grossman, P., Niemann, L., Schmidt, S., & Walach, H. (2004). Mindfulness-based stress reduction and health benefits: A meta-analysis. *Journal of Psychosomatic Research, 57*, 35-43. doi:10.1016/S0022-3999(03)00573-7

Helsdingen, A.S., van den Bosch, K., van Gog, T., & van Merriënboer, J.J.G. (2010). The effects of critical thinking instruction on training complex decision making. *Human Factors, 52*, 537-545. doi:10.1177/0018720810377069

Jha, A.P., Stanley, E.A., Kiyonaga, A., Wong, L., & Gelfand, L. (2010). Examining the protective effects of mindfulness training on working memory capacity and affective experience. *Emotion, 10*, 54-64. doi:10.1037/a0018438

Kabat-Zinn, J. (2003). Mindfulness-based interventions in context: Past, present, and future. *Clinical Psychology: Science and Practice, 10*, 144-156. doi:10.1093/clipsy/bpg016

Meland, A., Fonne, V., & Pensgaard, A.M. (2012). *Mindfulness based mental training in high performance aviation*. Paper presented at the Annual Meeting of Aerospace Medical Association, Atlanta, GA.

Orasanu, J. M., & Backer, P. (1996). Stress and military performance. In J. E. Driskell & E. Salas (Eds.), *Stress and human performance* (pp. 89–125). Mahwah, NJ: Erlbaum.

Passmore, J. (2009). *Mindfulness at Work and in Coaching.* Paper presented at the Danish Psychology Society Conference, Copenhagen, Denmark.

Ployhart, R.E., & Bliese, P.D. (2006). Individual adaptability (I-ADAPT) theory: Conceptualizing the antecedents, consequences, and measurement of individual differences in adaptability. In C.S. Burke, L.G. Pierce, and E. Salas (Eds.), *Understanding adaptability: A prerequisite for effective performance within complex environments* (pp. 3-39). Oxford, UK: Pergamon.

Rabeau, S. (2011). Formation des pilotes de transport tactique militaires [Training military tactical transport pilots]. *Proceedings of the Conference « Air Transport Pilots Facing the Unexpected »*, pp. 115-120.

Richardson, K.M., & Rothstein, H.R. (2008). Effects of occupational stress management intervention programs: A meta-analysis. *Journal of Occupational Health Psychology, 13*, 69-93. doi:10.1037/1076-8998.13.1.69

Schunn, C.D., & Reder, L.M. (1998). Strategy adaptivity and individual differences. *Psychology of Learning and Motivation, 38*, 115-154.

Stanley, E.A., Schaldach, J.M., Kiyonaga, A., & Jha, A.P. (2011). Mindfulness-based mind fitness training: A case study of a high-stress predeployment military cohort. *Cognitive and Behavioral Practice, 18*, 566-576. doi:10.1016/j.cbpra.2010.08.002

Stanovich, K.E. (2011). *Rationality and the reflective mind*. New York, NY: Oxford University Press.

Stanovich, K.E., & West, R.F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*, 672-695. doi:10.1037/0022-3514.94.4.672

Weick, K.E., & Sutcliffe, K.M. (2006). Mindfulness and the quality of organizational attention. *Organization Science, 17*, 514-524. doi:10.1287/orsc.1060.0196

The Expanded Cognitive Task Load Index (NASA-TLX) applied to Team Decision-Making in Emergency Preparedness Simulation

Denis A. Coelho[1], João N. O. Filipe[1], Mário Simões-Marques[2], Isabel L. Nunes[3,4]
[1]Universidade da Beira Interior, [2]Portuguese Navy,
[3]Universidade Nova de Lisboa, [4]UNIDEMI
Portugal

Abstract

The study demonstrates the use of the expanded TLX instrument (Helton, Funke & Knott, 2014) for cognitive and team-related workload self-assessment of 38 participants, solving the UNISDR – ONU stop disasters game simulation. Subjects in one group (GF; n=30) performed group decision-making without prior individual practice on the simulation. A subset of GF participants (n=6) subsequently reiterated the simulation alone, reassessing their cognitive workload. Another group (IF; n=8) individually performed the simulation and reiterated it in groups. Most GF participants, moving from group to singly conditions, reported decreasing physical and temporal demands, unchanged self-assessed performance, and increased mental demands, effort and frustration. IF participants incurred increasing mental, physical and temporal demands, as well as increased effort, with decreasing frustration and better performance, from singly to group conditions. Team workload results differed across groups; GF had higher levels of reported team dissatisfaction, equivalent assessments of team support and lower assessments of coordination and communication demands coupled with decreased time sharing as well as lower team effectiveness, compared to IF. Results bear implications on training of decision-making teams; singly training team members preceding group training supports team-decision making effectiveness and individual performance within teams going through first stages of a system learning curve.

Introduction

This section presents the interest in studying training for team-decision making and the scope of emergency preparedness. To this follows the presentation of the study aims, a methods section describing participants, the simulation and the experimental procedure, the results and their statistical analysis and, finally, a concluding discussion.

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

Training for team-decision making

Growing attention has been paid to the need to develop problem-specific models of problem solving, as opposed to traditional phase models articulating single approaches to solving all kinds of problems (Silber & Foshay, 2009). Work has become complex enough to require the use of teams at all hierarchical levels, with organizational success depending to a large extent on the ability of teams to collaborate and work effectively in solving complex problems (DeChurch & Mesmer-Magnus, 2010). Problem solving is also a learning process (Cooke et al. 2000) and team training benefits from a curriculum designed by a task analysis (Hamman, 2004). In the process of researching and understanding new information, the newly acquired understanding is added into the team's knowledge base, accumulating its experience from solving similar types of problems (Hung, 2013). According to DeChurch and Mesmer-Magnus (2010) relatively little is known about how team cognition forms and how to support it, dispite this being a critical issue for those designing teams and using teams in applied settings. The present study contributes to unveiling how to support the individual's performance within a decision-making team as well as team effectiveness.

This study investigates the effect of individual practice taking place prior to an otherwise unprepared group problem solving session (consisting of an emergency preparedness simulation) on individual and team-related workload. Studies focusing on workload measurement as a state should take a within-subjects perspective in their analysis (Helyton, Funke &Knott, 2014), although studies focusing on training evaluation often do not concurrently develop a within-subjects and a between-subjects perspective (Hagemann & Kluge, 2013). In this contribution, both within-subjects and between subjects perspectives are considered.

In this study, it is expected that the effect of training improves individual performance by the time of a second simulation run, irrespective of having done a first simulation run within a group or singly, or having done a second simulation run singly or within a group. This notwithstanding, it is expected at the onset of the study that first handedly and individually acquiring knowledge related to the problem at hand, prior to engaging in team-decision making within the process of solving the problem, will lead to improved team effectiveness. Individual practice following group interaction is used in the experiment as a means of balancing two group conditions, and enabling more extensive between subjects-analyses even if the primary interest of the study is supporting effective team- decision making.

Emergency preparedness and the nature of decision-making therein

Emergencies are unpredictable; needs for resources and information are difficult to define beforehand (Coelho, 2013-b). Emergency management is a mission that in several phases: work to avoid crises, preparation for crises, operative work, and evaluations after an event (Fig. 1).

Emergency management is a complex process requiring coordination of different actors, with different cultures, goals and views of the world. It aims to provide efficient and effective responses to multiple and often conflicting needs in situations

of scarce resources, considering several complementary functional elements, such as supply, maintenance, personnel, health, transport and construction. In all these elements the decision-making issues relate to basic questions: what, where, when, who, why, how, how much? These questions become particularly difficult to answer in critical situations, such as disaster relief, especially sensitive to the urgency and impact of decisions (Simões-Marques & Nunes, 2013). The commonly accepted phases of the management of the response to emergent events and critical disasters can be further characterized as follows: mitigation - preventing future emergencies or minimizing their effects, preparedness - preparing to handle an emergency, response - responding safely to an emergency, and, recovery - recovering from an emergency. The preparedness phase allows the development of an adequate level of resilience which enables effective emergency response and faster recovery, namely through a continuous cycle of planning and training (Fig. 2), as well as through public information, education and communication.

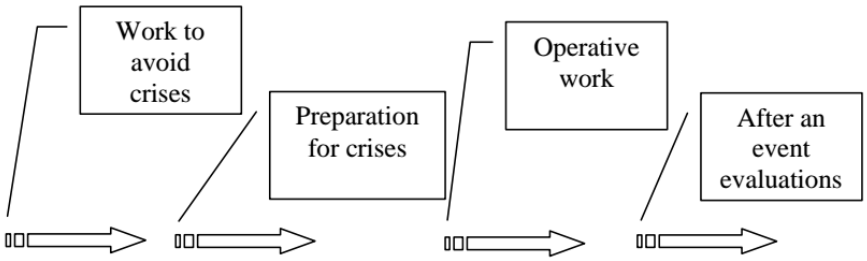


Figure 1. Phases in the management of the response to emerging events and critical disasters (Coelho, 2013-c).

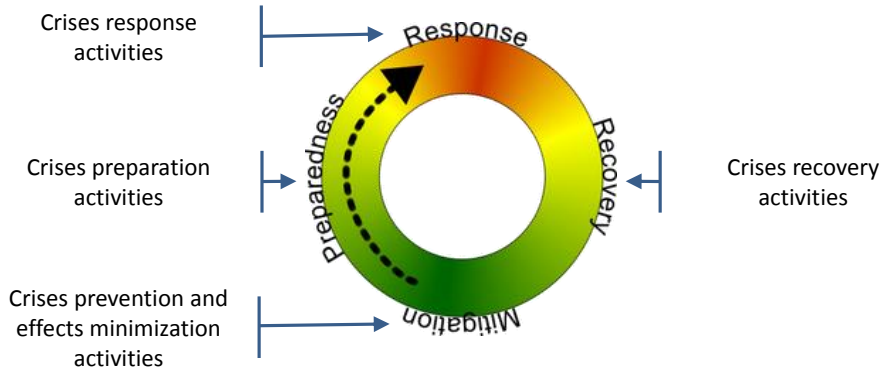


Figure 2. The continuous cycle of mitigation, preparedness, response and recovery.

According to Helton, Funke and Knott (2014) there is a growing interest in developing collaborative ways of teaching students about natural disasters (Berson & Berson, 2008; Gaillard & Pangilinan, 2010) as well as using simulation games to understand human behaviour in regard to disasters (Brigantic et al., 2009). The

simulation that is used in the experimental study deals with natural disaster preparedness, as a means of taking actions and altering the built environment as a way of mitigating the severity of the consequences of the disaster when it strikes, even if in reality it is uncertain when in the future it will occur.

Aims

Overall, this study is oriented towards empirically inducting knowledge contributing to support effectiveness of team decision-making and the individual's performance therein. The main aim of the experiment is to analyse the effect of individual problem-specific training on individual and team-related workload and performance/effectiveness in the course of a group decision-making activity.

Aditionally, an assumption was established in the design phase of the study. It was that practice leads to improved individual performance, irrespective of the order in which its two experimental conditions (group and solo) are experienced by the participant.

Method

Participants

Thirty-eight engineering students (13 women, 25 men), divided into two groups participated in the study for course credit. Their age ranged from 20 to 25 years. All study participants had normal or corrected-to-normal vision and hearing and none had any upper-body impairments limiting the use of a keyboard coupled with a computer pointing device (mouse) as interface. Participants were assigned to two groups. Table 1 presents participants count and sex by group, as well as subgroup size and gender mix.

*Table 1. Case counts for subgroup size and sex mix (legend: M - male sex; F - female sex; one of the subgroups in each category marked with * had 2 participants subsequently performing the simulation alone, for a total of 6 participants – 4 men and 2 women).*

Group	*Subgroup size*	*Quantity*	*Subgroup composition*		
			All male	*All female*	*Mixed*
GF – Group Simulation First (n=30; 8F; 22M)	2	2	1	—	1*
	3	3	1*	—	2*
	4	3	1	—	2
	5	1	—	—	1
IF – Individual Simul. First (n=8; 5F; 3M)	2	4	1	1	2

Simulation

The Stop Disasters game (www.stopdisastersgame.org) was developed by Playerthree© for the United Nations International Strategy for Disaster Reduction (UN/ISDR). In the Stop Disasters game (Fig. 3), players attempt to build disaster-resilient communities while also achieving development goals (e.g., building

infrastructure). In this study, we focused on an earthquake simulation, as it represents a regional interest for participants in Portugal. Because of course administration constraints, the time available for reiterations of the simulations was very limited (allowing only one to two per participant), which led to choosing the easiest setting. While most participants chose English, they were given the possibility of opting for the interface language that they felt most confident with of those available in the simulation game (English, Spanish or French). This game had previously been used for research (e.g. Khalid & Helander, 2013), but no team task analysis was available. The game yields a simulation performance score at the end of the simulation, which was not retained by the researchers.

Figure 3. Screenshot taken from the Stop Disasters Earthquake simulation game.

Expanded NASA TLX instrument for cognitive and team workload

NASA-TLX was established after an extensive three-year research effort and it sits properly in a web of correlations with external variables (Hart & Staveland, 1988). Workload has now become almost synonymous with the TLX (De Winter, 2014). Helton, Funke and Knott (2014) presented a modified version of the NASA-TLX that includes six additional team workload measures (Table 2). The additional team workload items were developed on the basis of literature review on teams carried out by Funke et al. (2012). The expanded version was used in this study in the decision-making in teams condition, while the standard version was used for the singly condition.

Procedure

The expanded version of the NASA-TLX instrument (Helton, Funke & Knott, 2014) was used to assess cognitive and team-related workload of a total of 38 students, divided into two groups (Fig. 4). Participants joined in teams of 2 to 5 people, solved the UNISDR – ONU stop disasters game simulation (earthquake challenge - easy mode) in a classroom setting. After the group simulation, each individual assessed his or her workload as well as the team-related workload using the expanded NASA-TLX. Subjects had no previous contact with the simulation and completed it within the allotted 25 minutes. A subset of 2 female and 4 male participants, who had made part of one of the two-person groups and of two of the three-person groups subsequently reiterated the simulation on their own, reassessing their cognitive task load, using the standard NASA-TLX.

Table 2.Rating scale definitions of the expanded (*) Task Load Index (TLX) (NASA, 1986, 2014; Helton, Funke & Knott, 2014) (these items were measured on 0-to-20 scales and multiplied by 5 to create comparable 0-to-100 scales).

Title	*Descriptions*
Mental Demand	How much mental and perceptual activity was required (e.g. thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration Level	How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?
*Coordination Demand	How much coordination activity was required (e.g., correction, adjustment)? Were the coordination demands to work as a team low or high, infrequent or frequent?
*Communica-tion Demand	How much communication activity was required (e.g. discussing, negotiating, sending and receiving messages)? Were the communication demands low or high, infrequent or frequent, simple or complex?
*Time Sha-ring Demand	How difficult was it to share and manage time between taskwork (work done as a team)? Was it easy or hard to manage individual tasks and those tasks requiring work with other team members?
*Team Effectiveness	How successful do you think the team was in working as a team? How satisfied were you with the team-related aspects of performance?
*Team Support	How difficult was it to provide and receive support (providing guidance, helping team members, providing instructions, etc.) from team members? Was it easy or hard to support/guide and receive support/guidance from other team members?
*Team Dis-satisfaction	How emotionally draining and irritating versus emotionally rewarding and satisfying was it to work as a team?

An unrelated group of 5 female and 3 male subjects individually performed the simulation (assessing their individual workload afterwards), and later, reiterated it in groups of 2 (assessing both their individual and team-related workload after the group simulation with the use of the expanded NASA-TLX). All assessments were made in the original language of the instrument. Statistical analysis was made with the assistance of IBM™ SPSS© 20 and using the approach described by Coelho et al. (2013-a).

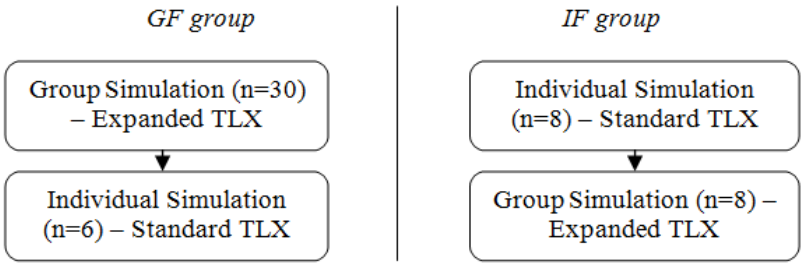


Figure 4. Diagram of experimental procedure.

Results and analysis

This section begins with the descriptive presentation of the results followed by their analysis (between subjects, within subjects and association of scales).

Presentation of results

Aggregated overall results are shown in Table 3, considering the condition that was rated and the order of the conditions in each group. The results overview suggests that within GF, effort and all types of demands increased for the participants involved in the two conditions, while performance and frustration remained almost unchanged. Conversely, for IF, performance increased and frustration decreased, while effort and all demands (mental, physical and temporal) increased. Looking across the team-related scales suggests higher coordination, communication and time sharing demands in the 2nd group, with much higher team effectiveness and equivalent team support. Selecting all participants in GF for comparison with IF, would suggest lower team dissatisfaction in IF, but the opposite ensues when selecting only the six participants in GF who reiterated the simulation alone.

Between subjects workload comparison (across both groups - group condition)

The independent samples Mann-Whitney test only yielded significant differences (significance threshold lowered to 0.001 to account for multiple comparisons – 12) across both complete groups in the group condition for communication demands ($U=10$; $p<0.001$) and for time sharing demands ($U=14$; $p<0.001$), both higher on average for IF. This would suggest that having more knowledge of the problem domain would require more communication and time sharing within the problem-solving setting in groups, even if groups are significantly smaller ($p<0.001$).

*Table 3. Mean and standard deviations obtained for each rating scale and group condition (legend: * - expansion team work related TLX rating scales; ** - subgroup of participants from GF who were subjected to the two experimental conditions).*

Rating scale	*GF (n=30)*			*IF*	
	Group 1st		*Solo 2nd*	*n=8*	
	n=30	***n=6*	***n=6*	*Group 2nd*	*Solo 1st*
Mental Demand	56 (19)	44 (27)	64 (12)	66 (18)	56 (23)
Physical Demand	36 (19)	25 (21)	33 (21)	48(26)	33 (21)
Temporal Demand	50 (16)	34 (17)	41 (11)	61 (26)	43 (18)
Performance	50 (22)	48 (28)	50 (28)	60 (30)	44 (31)
Effort	54 (20)	51 (29)	62 (23)	58 (29)	53 (18)
Frustration Level	52 (25)	40 (26)	42 (30)	45 (29)	64 (29)
*Coordination Demand	61 (19)	60 (25)		71 (16)	
*Communication Demand	64 (16)	68 (17)		94 (8)	
*Time Sharing Demand	54 (17)	48 (29)		88 (13)	
*Team Effectiveness	54 (20)	37 (23)		73 (17)	
*Team Support	64 (16)	67 (20)		66 (29)	
*Team Dissatisfaction	35 (22)	14 (18)		24 (22)	
Group size	3.6 (0.9)	2.7 (0.5)		2.0 (0.0)	

When selecting only the sub-set of participants in GF with smaller average team size, closer to the team size in IF, for comparison, more of the differences show significance, as the data summarised in the 2nd and the 4th columns of Table 3 are compared between each other. The differences that had been previously found when considering the whole GF are comfirmed for communication demands (U=3.5; p=0.00).

Association of scales (within subjects) for expanded instrument (both groups)

The 12 expanded NASA-TLX rating scales were correlated against each other yielding the significant results depicted in Table 4 (considering both groups below the diagonal and only GF above the diagonal, which may emphasize which associations are tied in part to differing experimental conditions and which are not; an association shown above and below the diagonal is deemed more robust). The positive moderate association between performance and mental demand shows up consistently in the top left quadrant of Table 4 (correlations amongst the standard TLX scales). Crossing the standard and expansion TLX rating scales shows that temporal demand is consistently positively correlated with team effectiveness and team dissatisfaction (but team effectiveness and team dissatisfaction do not correlate amongst each other). Within the new team workload scales, correlations are plentiful. Those significant and consistent below and above the diagonal of Table 4 lay between communication and coordination demands, as well as between team support and both communication and coordination demands. Team effectiveness

was found to be consistently moderately and positively correlated with both communication and time sharing demands.

Within subjects workload scale change (controlled for order of simulation type)

Aggregate change in each rating scale (the workload scale shown was obtained for each participant and condition by summing the ratings for mental, physical and temporal demands together with effort) is shown in Table 5. No statistical significance was found in the differences between the level of change that was incurred on the standard TLX and the compounded workload scales moving from the first simulation to the second one, across groups. Moreover, the one sample T-test, with test value zero, in GF, only showed significance (p=0.04) for mental demand change and workload change (p=0.02), while approaching significance (p=0.06) for effort change. In IF, tests did not yield significance.

The assumption that practice leads to improved individual performance, irrespective of the order in which its two experimental conditions (group and solo) are experienced by the participant, was further tested by joining both groups (last column in Table 5) and performing the one sample T-test for the test value of zero. This yielded significance for mental demands change (p=0.02), for physical demands change (p=0.03) and for workload change (p=0.02), but not for performance. Hence, the aforementioned assumption was not confirmed in the analysis.

*Table 4. Significant correlations (Spearman) encountered among the rating scales of the expanded TLX (legend: * - p< 0.05; ◊ - p< 0.01) joining both groups in the group condition (n=38) below the diagonal, and considering only GF above the diagonal (n=30).*

Rating scale	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1.Mental Demand	1			+.5*								
2.Physical Demand		1										+.4*
3.Temporal Demand	+.5◊		1							+.4*		+.4*
4.Performance	+.4*			1								
5.Effort					1							
6.Frustration Level						1						
7.Coordination Dem.						-.3*	1	+.5◊			+.6◊	
8.Communication D.							+.5◊	1		+.4*	+.5◊	
9.Time Sharing Dem.								+.6◊	1	+.4*		
10.Team Effectiven.			+.5◊				+.4*	+.5◊	+.4◊	1		
11.Team Support						-.3*	+.4*	+.4◊	+.3*		1	
12.Team Dissatisfact.			+.4*			+.4*		-.4*			-.4*	1

Table 5. Change in ratings of the standard TLX scales, from the first to the second simulation run, across groups (mean and standard deviation in parentheses; workload score obtained from adding effort to mental, physical and temporal demands ratings).

Standard TLX rating scale	*GF (n=6)*	*IF (n=8)*	*Both groups (n=14)*
Mental Demand (change)	20 (17)	11(24)	15 (21)
Physical Demand (change)	8 (19)	15 (20)	12 (19)
Temporal Demand (change)	7 (16)	18 (31)	13 (25)
Performance (change)	3 (30)	12 (51)	8 (42)
Effort (change)	11 (11)	4 (34)	7 (26)
Frustration Level (change)	2 (33)	-19 (46)	-10 (41)
Workload (change)	46(33)	48 (88)	47 (68)

Discussion

Effect of individual practice on group activity

Significant differences in the outcomes across two groups appeared for team communication and team time-sharing demands, which were higher for participants who had undergone singly practice prior to group activity. No significant differences were found across groups for individual performance and team effectiveness in the group condition.

Verification of assumption that practice leads to improved performance

Although on average there was an overall self-assessed performance increase of 8 percentage points (only 3% in GF and as much as 12% in IF) it was not significantly different from zero. Moreover, the conditions in GF may have increased the likelihood of a more intensified workload in the second simulation (carried out alone), for a marginal improvement in performance, compared to IF. Interestingly, workload (obtained from adding effort with mental, physical and temporal demands ratings) increased significantly from the first to the second experimental condition considering both groups united.

Conclusion

The results bear implications on training of decision-making teams, suggesting that singly practice of team members preceding group practice supports team-decision making effectiveness within teams going through the first stages of a system or problem-solving learning curve.

Limitations of the study

The study was based on a video-game based simulation. Kühn et al. (2014) reported on an anatomically based corroboration for association between frequent video-game playing and improvement in cognitive functions. Although participants had not previously interacted with the simulation used, previous experience with video-games at large was not controlled in this study. Hence, the evolution of each participant's individual workload and performance assessments from the first to the

second simulation run could have been influenced by general video-gaming experience.

References

Berson, I.R., & Berson, M.J. (2008). Weathering natural disasters with a net of safety. *Social Education*, *72*, 27–30.

Brigantic, R.T., Muller, G.A., Taylor, A.E., & Papatyi, A.F. (2009). Gaming to predict human responses to mass causality events. In *Proceedings of the International Conference on Computational Science and Engineering*, Vancouver: IEEE, 1194-1198.

Coelho, D.A., Harris-Adamson, S., Lima, T.M., Janowitz, I., & Rempel, D.M. (2013-a). Correlation between different hand force assessment methods from an epidemiological study. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *23*, 128-139.

Coelho, D.A. (2013-b). Cognitive engineering and emergency management. In *Proceedings of the 10th international conference on Engineering Psychology and Cognitive Ergonomics: understanding human cognition*-Volume Part I (pp. 197-204). Springer-Verlag.

Coelho, D.A. (2013-c). Editorial: Supporting cognition in the management of disasters and emergencies. *Int. J. Human Factors and Ergonomics*, *2*, 1-10.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. (2000). Measuring team knowledge. *Human Factors*, *42*, 151-173.

De Winter, J.C.F. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. *Cognition, Technology & Work*, 1-9.

DeChurch, L.A., & Mesmer-Magnus, J.R. (2010). The cognitive underpinnings of effective teamwork: a meta-analysis. *Journal of Applied Psychology*, *95*, 32-53.

Funke, G.J., Knott, B.A., Salas, E., Pavlas, D., & Strang, A.J. (2012). Conceptualization and measurement of team workload: A critical need. *Human Factors*, *54*, 36-51.

Gaillard, J.C., & Pangilinan, M.L. (2010). Participatory mapping for raising disaster risk awareness among the youth. *Journal of Contingencies and Crisis Management*, *18*, 175-179.

Hagemann, V., & Kluge, A. (2013). The effects of a scientifically-based team resource management intervention for fire service teams, *Int. J. Human Factors and Ergonomics*, *2*, 196-220.

Hamman, W. R. (2004). The complexity of team training: what we have learned from aviation and its applications to medicine. *Quality and Safety in Health Care*, *13(suppl 1)*, i72-i79.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: *Hancock, P.A., Meshkati, N. (eds.) Human mental workload*, North Holland Press, Amsterdam, 139-183.

Helton, W.S., Funke, G.J., & Knott, B.A. (2014). Measuring Workload in Collaborative Contexts: Trait Versus State Perspectives. *Human Factors*, *56*, 322-332.

Hung, W. (2013). Team-based complex problem solving: a collective cognition perspective. *Educational Technology Research and Development*, *61*, 365-384.

Khalid, H. M., & Helander, M. G. (2013). *Psycho-Cultural Analysis of Disaster Risk Attitudes in Situation Awareness*. Damai sciences sdn bhd, Kuala Lumpur (Malaysia).

Kühn, S., Lorenz, R., Banaschewski, T., Barker, G.J., Büchel, C., Conrod, P.J., Flor, H., Garavan, H., Ittermann, B., Loth, E., Mann, F., Nees, F. Artiges, E., Paus, T., Rietschel, M., Smolka, M.N., Ströhle, A., Walaszek, B., Schumann, G., Heinz, A., & Gallinat, J., The IMAGEN Consortium (2014). Positive association of video-game playing with left frontal cortical thickness in adolescents. *PloS one*, *9*, e91506.

NASA (1986). *Task Load Index (NASA-TLX). v. 1.0. Paper and pencil package (instruction manual)*. NASA Ames Research Center, Moffett Field, CA. http://humanfactors.arc.nasa.gov/groups/TLX/downloads/TLX.pdf

NASA (2014). *Task Load Index paper and pencil version*. NASA Ames Research Center, Moffett Field, CA. http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf

Silber, K.H., & Foshay, W.R. (eds.) (2009). *Handbook of improving performance in the workplace, instructional design and training delivery*. Vol. 1. John Wiley & Sons.

Simões-Marques, M., & Nunes, I.L. (2013). A Fuzzy multicriteria methodology to manage priorities and resource assignment in critical situations. In: *V. Zeimpekis, S. Ichoua, I. Minis (Eds.) Humanitarian & Relief Logistics: Research issues, case studies and future trends*, Operations Research/ Computer Science Interfaces Series (ORCS), Vol. 54, chapter 7, 129-153, Springer.

Evaluation of Crew Resource Management Interventions for Doctors-on-call

Vera Hagemann[1], Annette Kluge[2] & Clemens Kehren[3]
[1]University of Duisburg-Essen
[2]Ruhr University Bochum
[3]University Medical Centre Essen
Germany

Abstract

"Doctors-on-call" work in High Responsibility Teams, e.g. in hospitals or a (helicopter) emergency medical service (H/EMS), so called High Reliability Organisations. Due to their complex and demanding work contexts, where errors lead to severe consequences, doctors-on-call are required to develop non-technical competencies. To support reliable teamwork (aeromedical) crisis resource management (A/CRM) interventions have been implemented in initial training and further education more and more. The objective of this study is to evaluate the effectiveness of A/CRM interventions in initial training as well as in combined recurrent HEMS trainings for pilots, paramedics and doctors-on-call. Two interventions for doctors-on-call in initial training (n=79) and five interventions in HEMS training (n=71) were evaluated. Results of the pre-post-test-design for A/CRM for doctors-on-call initial training showed that the intervention was judged positively regarding usefulness and learning. Safety-relevant attitudes changed significantly ($.13 < \eta^2_p < .24$). The results for A/CRM in HEMS training also demonstrated effectiveness regarding usefulness and learning and safety-relevant attitudes increased significantly ($.28 < \eta^2_p < .41$). Due to a pre-post-post-test-design results showed stable attitude changes also three months later. So far, no studies exist documenting the valuable effects of A/CRM interventions for doctors-on-call in initial training and working in HEMS.

Introduction

Teams are a core element of a wide range of organisations. Given the increasing complexity of organizations and task fulfilment, teamwork is essential for success in meeting constantly changing requirements and reacting flexibly to turbulent business environments (Cannon-Bowers & Bowers, 2011; Hollenbeck et al., 2012). The advantage of teamwork is to use synergies of team members` competencies, knowledge and skills. Therefore, teams are able to adapt to changing conditions and cope with new situations successfully (Baker et al., 2006). Within some work environments the work has been structured as teamwork from the historical beginning of their professions, which means there was no period of time when it was en-vogue to implement teamwork with a special focus on teamwork processes, such as in the automobile industry or coal mines (cf. Hagemann, 2011, p.26). No one has

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

ever questioned the importance of teamwork within hospital anaesthesia teams, doctors-on-call or disaster management and first responder teams in the fire service. It would be barely conceivable that those people are not working as a team. But that does not mean that these teams have no teamwork problems.

Teams in healthcare, fire services, aviation or police units work in so-called High Reliability Organizations (HROs, Weick & Sutcliffe, 2003). They are named High Responsibility Teams (HRTs; cf. Hagemann et al., 2011) due to their dynamic and often unpredictable working conditions and demanding work contexts, in which technical faults and slips have severe consequences for human beings and the environment if they are not identified and resolved within the team immediately (Kluge et al., 2009). HRTs bear responsibility regarding their own lives and those of third parties based on their actions and consequences. In order to adapt to the dynamic und unpredictable working conditions successfully, they are confronted with specific requirements regarding information sharing and coordination – their non-technical skills (see e.g. Flin et al., 2005). Teamwork in HRTs is different from those in non-HRTs and is assumed to be very demanding (cf. Hagemann, 2011, pp.27-28). The impact on other peoples` life is enormous, especially when incidents or accidents occur. The notion that HRTs have always worked in teams does not imply that a particular team communicates and coordinates teamwork successfully. The human contribution to accidents and incidents in HRT-work has been recognised by many industries over the last three decades (Reason, 2008). The causal relationship between human error and teamwork problems such as breakdowns in communication or coordination processes or failures in decision-making and accidents and incidents was recognised. Examples of this are the Tenerife airport accident in 1977, which resulted in a loss of 583 lives, or the explosion of the Deep Water Horizon in spring 2010, which claimed 11 lives (Flin et al., 2002; Helmreich et al., 1999).

Teamwork professions such as medical teams in hospitals or doctors-on-call are as well recognizing the human contribution to errors and incidents. Examples in this regard include the tragic death of the (healthy) 2-month-old Jose Martinez in a hospital in Houston in 1996 due to medication errors (Belkin, 1997) or the death of the cardiac Rosemarie Voser who received a donor heart with a wrong blood type in Zurich due to a misunderstanding in communication[4]. It is estimated that about 44.000 up to 98.000 people in the USA die each year as a result of medical errors (Kohn et al., 2007). These examples show that HRTs also need support in their teamwork processes and special attention from teamwork experts, even though they work in teams for a very long time. The aim of the present studies is thus to explore the positive effects of a special kind of team training for *doctors-on-call* in hospital teams or helicopter emergency medical service (HEMS) teams on team members' *reactions* and teamwork safety-relevant *attitudes*.

[4] http://www.news.ch/Fehler+bei+Herz+OP/214105/detail.htm

Doctors-on-call's work (environment)

There are two different models of the pre-hospital medical care in emergency cases. On one hand the "Anglo-American model" which operates with pre-hospital care specialists, such as paramedics or emergency medical technicians (EMTs). Doctors-on-call (also called emergency physicians) are not part of this model in the pre-hospital care. In contrast the so called "Franco-German-Model" which is led by physicians and supported by paramedics. This is also typical for most countries in Central Europe. Doctors-on-call in Germany provide the necessary medical interventions for patients in emergency medical service. Therefore they are active as well in road ambulances as in helicopter emergency medical services. Germany is one of the rare countries in the world having established a complete comprehensive network of helicopter emergencies. Hence, every patient can be reached within 15 minutes by a rescue helicopter from its more than 70 bases. Some of the rescue helicopters provide a day and night service, 24/7. Although the rescue network is comprehensive in Germany, all team members of the emergency medical technicians still face the challenge of reaching the emergency scene within minutes.

Regularly, confrontations with seriously injured patients, e.g. after motor vehicle crashes, but also with victims of crime scenes or outbreak of violence occurs. Due to these particular cognitive and social demands a close co-operation with the police, beside the collaboration with the fire department or the emergency rooms in hospitals is required. To take live care / life support decisions within seconds in the rescue unit during the assignment, teamwork is the key.

Crew or Aeromedical Crisis Resource Management

Professions such as surgery, anaesthesia, or doctors-on-call are recognizing the human contribution to errors and incidents and are trying to help themselves by applying a team training intervention originally developed for aviation personnel in order to accomplish the challenges of their demanding and complex teamwork contexts (cf. Gaba et al., 2001; Müller et al., 2007). This intervention, called Crew Resource Management (CRM) was developed to improve teamwork-relevant non-technical skills (e.g. communication or adaptation) of team members and increase team effectiveness and safety in HRTs. CRM has been defined as "the use of all available resources to achieve safe and efficient flight operations" (Lauber, 1984, p. 20). CRM-based training concepts are instructional strategies for HRTs in order to a) train them to use *all available resources* efficiently (i.e. people, equipment, and information), b) enhance their *teamwork* and therefore enhance their *performance*, and c) diminish the likelihood of possible *human error* with severe consequences for people and the environment (Salas et al., 2006a).

In its early stages, CRM mainly focused on pilots. During the 1990s, it was extended to flight attendants and maintenance technicians (Helmreich et al., 1999) and today it is also prescribed by law for all aviation personnel worldwide (EU OPS 1). CRM-based training concepts have been well established within commercial aviation for over 30 years. During this time span and due to this training concept, that focuses on team members' non-technical skills and error management, incident and accident rates have been reduced (Flin et al., 2002). At the end of the 1990s, a tendency to

apply CRM within anaesthesia could be observed. Specific team training interventions called aeromedical or anaesthesia crisis resource management (ACRM) were developed (Davies, 2001; Gaba et al., 2001). Since then, more and more HRTs in the fire service or surgery are trying to transfer CRM from aviation to their own teams, called, for instance team resource management for the fire service (cf. Hagemann & Kluge, 2013; Okray & Lubnau, 2004).

Some meta-analyses support the effectiveness of CRM interventions on teamwork relevant competence acquisition in HRTs for aviation and military or medical teams as well as in the oil industry. For example, Salas et al. (2006b) report in their meta-analysis—100 studies included—positive effects of CRM on team members' reactions and teamwork safety-relevant attitudes. Diverse results (positive or no effects) are reported in regard to teamwork safety-relevant knowledge and behaviour as well as on organisational outcomes. The meta-analysis conducted by O`Connor et al. (2008) included 16 studies and supports positive training effects. The reported studies demonstrated positive effects of CRM on team members' reactions, teamwork safety-relevant attitudes and behaviour. In regard to a safety-relevant knowledge gain medium effects were found.

The positive influence of teamwork relevant competencies and accordingly team processes on team performance has also been demonstrated in some studies. Schmutz and Manser (2013) included 28 studies in their review and report medium to large effect sizes regarding the positive effects of team process behaviours on clinical performance, such as task management, problems during operation, operating time, or morbidity. Because A/CRM interventions support teamwork relevant competence acquisition and teamwork competencies influence clinical performance positively, A/CRM is a very powerful "instrument" in supporting HRTs for reliable teamwork, also in a medical context.

So far, nearly nothing is known about the effects of A/CRM interventions on prerequisites for successful teamwork of doctors-on-call, working in e.g. hospitals or HEMS. In order to assess positive training effects on teamwork relevant competence acquisition for this target group, the widely used training evaluation hierarchy from Kirkpatrick (1998) is applied. This hierarchy categorises training outcomes on four levels. The first two levels are considered here. The first level is the evaluation of "reactions", such as subjectively perceived enjoyment and perceived usefulness of the A/CRM intervention. The second level is "learning" and contains the participant's attitudinal changes and knowledge gain after A/CRM intervention. "Behavioural changes" are the hierarchy's third level and refer to the application of acquired knowledge and skills to the job. This level will be considered indirectly in the evaluation based on questions regarding the transfer climate. The aim of the present paper is to demonstrate the positive effects of A/CRM interventions for doctors-on-call on team members' *reactions, subjectively rated learning success,* and *teamwork safety-relevant attitudes.*

Hypotheses

As demonstrated in the meta-analyses by Salas et al. (2006b) and O`Connor et al. (2008), *first*, it is assumed that the A/CRM interventions will have a positive impact on team members' *reactions*.

Second, it is assumed that the A/CRM interventions will have a positive impact on team members' subjectively rated *learning success* (knowledge and attitude).

Third, it is assumed that the A/CRM interventions will have a positive impact on the *teamwork-relevant attitudes* such as those demonstrated by Gregorich, Helmreich, and Wilhelm (1990) and Helmreich and Wilhelm (1991). Evaluating this effect is important, because in training research it is assumed that positive changes in attitudes (affective levels) are one essential prerequisite for changes in (safety-relevant) behaviour (O'Connor, Flin, Fletcher & Hemsley, 2003).

Furthermore, it is analysed whether doctors-on-call have the possibilities to apply the newly learned concepts and skills in training during their daily work or not. That means the transfer climate is measured.

Method

Samples

Two samples were included in this study. One sample consisted of doctors-on-call in initial training (n = 79). These doctors have been in vocational training to become doctors-on-call. The sample consisted of two subgroups which got the initial training at different times. 38 doctors were male, 31 female, and 10 doctors did not indicate their sex. Their mean age was M = 32.63 years (SD = 7.40). 13.5% of the doctors declared that they already had participated in any sort of A/CRM intervention before.

The second sample consisted of doctors-on-call, pilots and paramedics (helicopter crew member, HCM) in combined recurrent HEMS trainings (n = 71). Five groups in combined recurrent HEMS training that received an A/CRM intervention were includes in the study. 60 people were male, 5 female, and 6 people did not indicate their sex. Their mean age was M = 42.1 years (SD = 7.45). 25 people were pilots, 21 were HCMs and 23 were doctors-on-call (2 missing). 32 people (47.1%) declared that they already had participated in any sort of A/CRM intervention before. 23 people of those 32 were pilots, 6 HCMs and 3 were doctors-on-call. The mean age for doctors-on-call was M = 42.4 years (SD = 5.8), 19 of them were male and 4 female.

Field study design

The study included two within-group comparisons with a pre-post-test design for doctors-on-call in initial training (sample 1, S1) and a pre-post-post-test design for participants in combined recurrent HEMS training (sample 2, S2). Due to organisational constraints and patient needs the participants were not able to visit the

interventions all at the same time. Hence, the A/CRM interventions were conducted two (S1) respectively 5 (S2) times in the same manner.

The doctors-on-call in S1 participated in an A/CRM seminar of one and a half hours duration. This seminar was integrated into a one week vocational training in a German hospital to become a doctor-on-call. Other seminar topics for example were trauma support, cardio-pulmonal-resuscitation, specifics of the emergency medical system EMS, and cooperation with fire brigade, HEMS, Search and Rescue SAR, paramedics / EMTs emergency medical technicians. The instructor of the A/CRM seminar came from an aviation and medical background. The discussed topics were human factors, error management, communication, and situation awareness and its influences on human behaviour and teamwork. The design of the seminar consisted of theoretical inputs and discussion phases. Seminars based on such a design are able to influence reaction, attitudes and knowledge, the first two levels of Kirkpatrick's evaluation hierarchy (1998).

Table 1. Overview of the study design for sample 1 and 2

Sample 1	*One day before the A/CRM seminar*		*At the end of the seminar day*	
	T0		*T1*	
Instruments	Attitudes	A/CRM seminar (90 minutes)	Attitudes Reactions to the A/CRM seminar Subjectively rated learning success	
Sample 2	*At the beginning of the first training day*		*At the end of the last training day*	*Three months later*
	T0		*T1*	*T2*
Instruments	Attitudes	A/CRM training (3 days)	Attitudes Reactions to the A/CRM training Subjectively rated learning success	Attitudes Transfer Climate

The participants in S2 got an A/CRM training of three days duration. The training was a combined training for pilots, HCMs, and doctors-on-call all working in a helicopter emergency medical service in Germany. This combination of participants is due to the fact that a helicopter crew in missions consists of one pilot, one HCM, and one doctor-on-call. The underlying proposition is that the people who work together should also be trained together. The discussed topics were human factors, error chains, attitudes, communication and coordination, leadership, situation awareness, and shared mental models and its influences on human behaviour and teamwork processes and outcomes. The design of the training was interactive and

consisted of a mixture of theoretical inputs, exercises, discussions, and reflections. Trainings based on such a design are able to influence attitudes, knowledge and behaviour, the first three levels of Kirkpatrick's evaluation hierarchy.

In S2 also the team members' behaviour was influenced. Because it was not possible to directly assess the behaviour of the team members after training all participants from S2 were asked to fill in a questionnaire measuring the transfer climate within their daily work three months after the training. The aim was to analyse whether the participants have the possibilities to apply the newly learned concepts and skills in training during their daily work.

The instruments measuring the team members' reactions to the A/CRM interventions, the subjectively rated learning success, and the teamwork safety-relevant attitudes were distributed in S1 one day before the seminar (T0) and at the end of the seminar day (T1). In addition to the listed instruments here S2 also worked on an instrument measuring the transfer climate. The instruments were handed out at the beginning of the first day of training (T0), at the end of the last day of training (T1) and three months later (T2) (see table 1). Due to this long time span and because of holidays, shift changes, and absenteeism, not all doctors-on-call, pilots, and HCMs were able to participate at all three measurement times.

Applied measuring instruments

Teamwork safety-relevant attitudes
To measure a change in teamwork safety-relevant attitudes an adapted version of the Fire Service Management Attitudes Questionnaire (FSMAQ, Hagemann, 2011) was applied two (T0, T1) or three times (T0, T1, T2) for sample 1 and 2, respectively. The questionnaire consisted of 20 items (five-point Likert scale from 1 to 5) and is called Doctors-on-call Management Attitudes Questionnaire (DMAQ). Other well established instruments have been the basis for this attitude questionnaire, e.g. ORMAQ surgery (Yule et al., 2004), ORMAQ anaesthesia (Sexton et al., 2000), CMAQ cockpit (Gregorich et al., 1990), and CAQ (McDonald & Shadow, 2003). The questionnaire covered the eight most frequently investigated safety-relevant attitudes: *command roles and responsibilities* (4 items, e.g., 'Team members should not question the decisions or actions of senior staff'), *speak up* (2 items, e.g., 'I inform other team members when my workload is too high'), *debriefing* (2 items, e.g., 'A regular debriefing of procedures and decisions after a mission is an important part of teamwork'), *feedback and critique* (2 items, e.g., 'Disagreements in the team are appropriately resolved, i.e., it is not 'who' is right, but what is best for the mission'), *realistic appraisal of stress* (3 items, e.g., 'Personal problems can adversely affect my performance'), *denial of stress* (3 items, e.g., 'A professional doctor-on-call is able to hide personal problems during the whole mission'), *handling errors* (2 items, e.g., 'I am more likely to make errors in tense or hostile situations'), and *teamwork* (2 items, e.g., 'I enjoy working in a team').

Subjectively perceived training outcomes
The training evaluation inventory (TEI; Ritzmann et al., 2014; Hagemann & Kluge, 2014) was applied for evaluating the A/CRM interventions and the team members' reactions and subjectively rated learning success, respectively, at T1. This inventory

consisted of 16 items (five-point Likert scale from 1 to 5). It covered training outcomes based on the first (*reaction*) and second (*learning*) level of Kirkpatrick's (1998) four levels of evaluation. Based on the work of Alliger et al. (1997), Phillips and Phillips (2001) and Salas et al. (2006a), the first level (reaction) was further divided into three scales: *reported enjoyment* (3 items, e.g., 'I enjoyed learning'), *perceived difficulty* (3 items, e.g., 'I understood all technical terms') and *perceived usefulness* (4 items, e.g., 'The training is useful for my profession'). In particular, perceived usefulness is assumed to support the motivation to apply acquired knowledge and skills to the trainees' field of work (Helmreich & Wilhelm, 1991; Phillips & Phillips, 2001; Salas et al, 2006c). Furthermore, it enhances the probability of the trainees' work performance improving. The second level (learning) was divided into *learning knowledge* (3 items, subjectively rated learning success, e.g., 'I think my knowledge has been expanded in the long term') and *learning attitudes* (3 items, e.g., 'I would recommend this training to my colleagues'). The subjectively rated learning success proved to be a successful predictor in relation to objectively measured learning success or knowledge acquisition (Ritzmann et al., 2014), and was therefore used as in indicator for the second level of Kirkpatrick's evaluation hierarchy. The TEI was used as it was developed for training evaluation and has been applied in various CRM training evaluation studies (see Ritzmann et al., 2014).

Transfer climate

In order to analyse whether doctors-on-call have the possibilities to apply the newly learned concepts and skills in training during their daily work a transfer climate questionnaire was applied at T2 (only S2). The instrument consisted of 15 items (five-point Likert scale from 1 to 5) and was developed based on the transfer climate questionnaire by Thayer and Teachout (1995). The questionnaire covered cues, reinforcements, and extinction possibilities. The scales were *goal cues* (3 items, e.g. 'My supervisors set performance goals that encourage me to apply the skills learned in the ACRM-training'), *social cues* (2 items, e.g. 'My colleagues help me applying the concepts learned in the ACRM-training at work'), *task cues* (2 items, e.g. 'We have the resources (equipment, human power, time) in order to fulfil the work as learned in the ACRM-training'), *positive reinforcement* (3 items, e.g. 'My supervisors appreciate it when I transfer the things learned in the ACRM-training to work'), *negative reinforcement* (3 items, e.g. '(Experienced) Colleagues make fun of the concepts communicated in the ACRM-training'), and *extinction* (2 items, e.g. 'I have only a few possibilities to apply the skills learned in the ACRM-training, so it is difficult for me to internalise them').

Results

In the following the three hypotheses will be tested for sample 1 (doctors-on-call initial training) and sample 2 (combined recurrent HEMS training). The last research question regarding the transfer climate will be tested for sample 2 only.

Table 2. M, SD, and Cronbach's α of training outcome scales at T1 for sample 1 and 2

	A/CRM seminar (S1), n = 79			*A/CRM training (S2), n = 71*		
Scales	*α*	*M*	*SD*	*A*	*M*	*SD*
Reported Enjoyment	.88	4.05	0.76	.80	4.65	0.52
Perceived Usefulness	.87	4.23	0.74	.84	4.70	0.43
Perceived Difficulty[+]	.82	4.47	0.59	.60	4.54	0.43
Learning Knowledge	.81	3.92	0.78	.81	4.25	0.58
Learning Attitudes	.92	4.21	0.83	.82	4.75	0.41

Notes. [+] A high score means that the training was not difficult; range from 1 to 5

In order to test hypotheses 1 and 2, the *subjectively perceived outcomes of the A/CRM seminar* and *the A/CRM training* were evaluated by applying the TEI at T1. The internal consistencies and means of the evaluation scales regarding reaction and learning are displayed in Table 2 for both samples. The mean values of the five scales indicated an overall very positive evaluation of the seminar or rather the training regarding team members' *reactions* and subjectively rated *learning success* in both samples. Hence, the results supported Hypotheses 1 and 2.

In order to test the third hypothesis, whether the *teamwork safety-relevant attitudes* changed positively and significantly after the A/CRM seminar/training, univariate analyses of variance (ANOVA) with repeated measures for analysing within-group effects were conducted – for each of the eight scales. The attitudes at T0 and T1 were within-subject factors.

Table 3. M, SD (in brackets), α and results of ANOVA with repeated measures regarding attitudes at T0 compared to T1 for sample 1 (n = 60) and 2 (n = 65)

Sample 1	*α*	T0	T1	*F*	*P*	η^2_p
Command roles and responsibilities	.72	4.39 (0.55)	4.43 (0.61)	0.663	.419	.01
Speaking up	.56	3.75 (0.72)	4.05 (0.81)	10.585	.002	.15
Debriefing	.37	4.78 (0.36)	4.78 (0.38)	0.000	1.00	.00
Feedback and critique	.21	4.01 (0.67)	4.02 (0.69)	0.012	.912	.00
Realistic appraisal of stress	.67	3.97 (0.73)	4.18 (0.67)	11.457	.001	.16
Denial of stress[+]	.70	2.98 (0.89)	2.63 (0.98)	18.903	.000	.24
Handling errors	.69	3.79 (0.85)	4.07 (0.82)	8.460	.005	.13
Teamwork	.28	3.90 (0.61)	3.98 (0.65)	1.065	.306	.02
Sample 2	*α*	*T0*	*T1*	*F*	*P*	η^2_p
Command roles and responsibilities	.63	4.26 (0.46)	4.50 (0.41)	26.024	.001	.30
Speaking up	.68	3.78	4.05	10.286	.002	.14

		(0.70)	(0.59)			
Debriefing	.35	4.35 (0.52)	4.58 (0.52)	10.091	.002	.14
Feedback and critique	.74	3.78 (0.56)	3.95 (0.65)	4.089	.050	.06
Realistic appraisal of stress	.88	4.04 (0.61)	4.32 (0.71)	21.351	.001	.25
Denial of stress[+]	.83	2.84 (0.80)	2.35 (0.79)	32.699	.001	.34
Handling errors	.54	3.67 (0.65)	3.97 (0.84)	8.727	.004	.12
Teamwork	.30	4.26 (0.59)	4.34 (0.58)	1.438	.235	.02

Notes. [+] Low values indicate a positive attitude; range from 1 to 5.

Referring to S1, there were no significant results for "command roles and responsibilities", "debriefing", "feedback and critique", and "teamwork" (see Table 3). Regarding "speaking up" ($F_{(1/60)} = 10.585$, $p < .01$, $\eta^2_p = .15$), "realistic appraisal of stress" ($F_{(1/60)} = 11.457$, $p < .01$, $\eta^2_p = .16$), "denial of stress" ($F_{(1/60)}$ 18.903, $p < .001$, $\eta^2_p = .24$), and "handling errors" ($F_{(1/60)} = 8.460$, $p = .001$, $\eta^2_p = .13$) the main effects for measurement time reached significance and the effect sizes were medium to large. Thus, these four attitudes changed significantly and positively from T0 to T1. The doctors-on-call showed a significant positive change in speaking up, realistic appraisal of stress, denial of stress, and handling errors.

Referring to S2, there were no significant results for "feedback and critique" and "teamwork" (see Table 3). Regarding "command roles and responsibilities" ($F_{(1/63)} = 26.024$, $p < .001$, $\eta^2_p = .30$), "speaking up" ($F_{(1/65)} = 10.286$, $p < .002$, $\eta^2_p = .14$), "debriefing" ($F_{(1/65)} = 10.091$, $p < .002$, $\eta^2_p = .14$), "realistic appraisal of stress" ($F_{(1/65)} = 21.351$, $p < .001$, $\eta^2_p = .25$), "denial of stress" ($F_{(1/65)} = 32.699$, $p < .001$, $\eta^2_p = .34$), and "handling errors" ($F_{(1/65)} = 8.727$, $p < .004$, $\eta^2_p = .12$) the main effects for measurement time reached significance and the effect sizes were all medium to large. Hence, these six attitudes changed significantly and positively from T0 to T1. Summing up, hypothesis 3 could be partially supported; both, the A/CRM seminar and training had a positive impact on safety-relevant attitudes.

In order to test whether the six *attitude* changes from T0 to T1 were stable over a time period of three months, paired samples t-tests were calculated to compare the results between T1 and T2 (see Table 4). This analysis was conducted for S2 only, because only this sample had a follow-up evaluation three months later. The six attitudes—command roles and responsibilities, speaking up, debriefing, realistic appraisal of stress, denial of stress, and handling errors—remained stable over time, as no difference between T1 and T2 reached significance (two-tailed). Summing up, the results indicate that the six positive attitude changes from T0 to T1 were stable over a period of three months.

Table 4. Means and results of paired samples t-tests of four attitudes between T1 and T2 (which changed significantly from T0 to T1) (n = 12)

	T1	T2	*T*	Sig. (two-tailed)
Command roles and responsibilities	4.55	4.63	$t(11) = -1.05$	$p > .32$
Speaking up	4.21	4.38	$t(11) = -1.00$	$p > .34$
Debriefing	4.70	4.63	$t(11) = 1.00$	$p > .34$
Realistic appraisal of stress	4.33	4.19	$t(11) = .68$	$p > .51$
Denial of stress[+]	2.42	2.53	$t(11) = -.51$	$p > .62$
Handling Errors	4.21	3.67	$t(11) = 2.24$	$p = .05$

Notes. [+] Low values indicate a positive attitude; range from 1 to 5.

Furthermore, descriptive data were analysed in order to answer the last research question, whether the pilots, HCMs, and doctors-on-call had the possibilities to apply the newly learned concepts and skills in training during their daily work. For this purpose a *transfer climate* questionnaire was applied in S2 at T2. The internal consistencies and means of the evaluation scales regarding cues, reinforcements and extinction are displayed in Table 5. The mean values of the six scales indicated an overall very positive transfer climate at work for the participants. Thus, the results indicated good possibilities to apply newly learned skills in training at work.

Table 5. M, SD, and Cronbach's α of transfer climate scales at T2 for sample 2

	A/CRM training (S2), n = 12		
Scales	*α*	*M*	*SD*
Goal Cues	.87	3.86	0.96
Social Cues	.91	3.67	0.98
Task Cues	.26	3.75	0.66
Positive Reinforcement	.81	3.92	0.95
Negative Reinforcement[+]	.59	3.00	0.90
Extinction[+]	.78	3.54	1.05

Notes. [+] High values indicate a positive transfer climate; range from 1 to 5.

All results were controlled for age and sex differences. No impacts of age and sex on the effects could be detected.

Discussion

The goal of the present studies was to investigate the positive impact of A/CRM seminars and trainings on doctors-on-calls' *reactions, subjectively rated learning success,* and *teamwork safety-relevant attitudes.* The first two hypotheses were supported; the third one was partially supported. According to the first hypothesis, the team members—in both samples—reported that they enjoyed the A/CRM seminar/training and that it was easy for them to follow it. They perceived the seminar/training as useful for their work and stated that they would, for example, recommend it to their colleagues. According to the second hypothesis, they developed a positive attitude towards the seminar/training and teamwork-relevant topics, respectively, and stated that they learned a lot. These results confirm findings

of previous evaluation studies of CRM training within aviation, military and fire service as demonstrated by Hagemann and Kluge (2014), Ritzmann et al. (2011), Salas et al. (1999), Salas et al. (2001), and Salas et al. (2006a). The present findings seem to indicate that A/CRM seminars and trainings for doctors-on-call could also be useful for enhancing non-technical teamwork competencies.

Evaluating *perceived usefulness* of a seminar or training is also important, because studies showed positive relationships between perceived usefulness of an intervention and transfer motivation, subjectively rated learning success as well as objective measurement of knowledge acquisition and maintenance (Alliger et al., 1997; Hagemann & Kluge, 2013, 2014; Helmreich & Wilhelm, 1991). Furthermore, the *subjectively rated learning success* is also a reliable predictor for objective measurement of knowledge acquisition (Ritzmann et al., 2014). The reported results indicate, that the doctors-on-call expended their knowledge regarding safety relevant teamwork competencies. These findings stress the importance of evaluating trainee reactions in a differentiated manner by focusing on perceived usefulness and subjectively rated learning success.

According to hypothesis three, positive changes in *teamwork safety-relevant attitudes* could be found in both samples. Results for the first sample showed that after the A/CRM seminar, four of the eight attitudes changed. These four were "speaking up", "realistic appraisal of stress", "denial of stress", and "handling errors". Results for the second sample showed that after the A/CRM training, six of the eight attitudes changed. These six were "command roles and responsibilities", "speaking up", "debriefing", "realistic appraisal of stress", "denial of stress", and "handling errors". Possible explanations for why attitudes regarding "realistic appraisal of stress" and "denial of stress" changed significantly in both samples might be that the seminar/training focused these topics deeply. To pick the link between handling and denial of stress—factors which influence performance negatively—and accidents and incidents out as a central theme is very common in A/CRM interventions. In sample 2, but not in sample 1, after the training the participants showed more positive attitudes regarding "command roles and responsibilities" and "debriefing". These differences could be explained by the thematic setting of priorities. In sample 1 the doctors-on-call were 10 years younger on average and at the beginning of their career as a doctor-on-call than the team members in sample 2. Hence, different teamwork relevant competencies are important for these target groups. More experienced team members might be more interested in leadership topics and instruments to steer team processes, such as debriefings.

These positive attitude changes are in accordance with some previous studies within other HRTs in aviation (Gregorich et al., 1990; Helmreich & Wilhelm, 1991) or in fire service teams (Hagemann & Kluge, 2013). Even though Röttger et al. (2013) did not report effects of CRM training on attitude changes, however they found significant relationships between negative attitudes and teamwork behaviour and performance in the maritime domain. In the present study it was also demonstrated that after a time period of three months, the positive attitude changes were stable (only S2). The demonstrated positive attitude changes are an important prerequisite

for showing safety-relevant behaviour during missions (Sexton & Klinect, 2001). Furthermore, the attitude changes demonstrated in the present studies are not common findings. O'Connor et al. (2012, p. 30) report, that many of the studies examining the impact of CRM training on attitude changes did not find any significant effects. Moreover, psychometric properties of the applied instruments were lacking. In their own study with naval aviators, O'Connor et al. also did not find any significant effects of CRM training on attitude changes. They report the psychometric properties of their inventory, which ranged from α = .44 to α = .59. These internal consistencies were typical of this type of questionnaire. The internal consistencies of the DMAQ within the present studies are in line with these results predominantly and ranged from α = .21 to α = .88.

The last research question focused on the *transfer climate* in sample 2. It was of interest, whether participants have the possibilities to apply the newly learned concepts and skills in training during their daily work or not. Because it was not possible to directly assess the behaviour of the team members during work after training they were asked to fill in a questionnaire measuring the transfer climate within their daily work three months after the training. The underlying idea was that the A/CRM trainings are able to influence not only knowledge and attitudes, but also behaviour. But new behaviour congruent to training will not or hardly be shown if there is no transfer climate. So transfer climate is a prerequisite to experience newly acquired behaviour (Greif & Kluge, 2004; Thayer & Teachout, 1995). The results of the present study show, that the doctors-on-call, the pilots, and the HCMs reported a good transfer climate after training. The aspect regarding positive reinforcement was assessed most positively.

Summing up, the findings indicate that A/CRM interventions for doctors-on-call are useful in terms of enhancing non-technical teamwork competencies, especially reactions, learning, and attitudes, but also behaviour. Furthermore, other research, for example regarding Bridge Resource Management training for navy teams (cf. O'Connor, 2011), indicates that not all kinds of CRM adaptations successfully lead to positive training outcomes, and indeed some have no effect at all. Thus, the findings of the present studies broaden the field regarding effective applications of A/CRM interventions. As a result medical services should consider implementing ACRM into their education and further trainings. A/CRM should be implemented into the curricula equal to other topics, not only for medical students but also for doctors in initial training to become doctors-on-call or in further education.

Limitations and Outlook

With regard to methodological problems, the DMAQ for evaluating the teamwork safety-relevant attitudes showed problems regarding reliability aspects. Some scales (e.g. teamwork or debriefing) had very low internal consistencies. These problems regarding attitude evaluations are common in the scientific community, as discussed earlier, but further research is needed for developing reliable and valid instruments for assessing attitudes.

Team diversity was not taken into consideration neither in the present studies nor in the A/CRM seminar or training. Jackson and Joshi (2011) stated in their review that

work team diversity "is likely to impede frequent and effective communication among team members" (p. 661) and has diverse—positive as well as negative—influences on team performance (p. 666). Hence, possible effects of team diversity on team performance should be taken into consideration in future studies. Furthermore, the topic "team diversity" and its implications on team performance and team processes should be implemented into A/CRM interventions. Today, medical teams become more and more diverse regarding gender, age, nationality, personality, attitudes, values, educational level or organizational tenure.

The third level (behaviour) in Kirkpatrick's (1998) evaluation hierarchy was assessed indirectly; the fourth level (outcomes) was not assessed at all in the present studies. This is a well known phenomenon in training evaluation studies. It costs a lot of time and resources to do that, but for further research it is required to evaluate behaviour at work after training. Also objective measures or so called hard facts (e.g. no complication during surgery or patient alive) should be analysed in order to assess training outcomes, which means the effects of A/CRM interventions on team performance as defined by patient well-being.

Summing up, the studies indicate the usefulness of A/CRM interventions for doctors-on-call on their non-technical teamwork competencies, even if the people do not have any prior experience with this kind of intervention. The foundations for more research regarding A/CRM interventions for doctors-on-call are led.

References

Alliger, G.M., Tannenbaum, S.I., Bennett, W., Traver, H., & Shotland, A. (1997). A Meta-Analysis of the Relations among Training Criteria. *Personnel Psychology, 50*, 341-358.

Baker, D.P., Day, R., & Salas, E. (2006). Teamwork as an Essential Component of High-Reliability Organizations. *Health Services Research, 41*, 1576-1598.

Belkin, L. (1997). How can we save the next victim? *The New York Times.*

Cannon-Bowers, J.A., & Bowers, C.A. (2011). Team Development and Functioning. In S. Zehdeck (Ed.), *APA Handbook of Industrial and Organizational Psychology Building and Developing the Organization* (Vol. 1, pp. 597-650). Washington: American Psychological Association.

Davies, J.M. (2001). Medical applications of crew resource management. In E. Salas, C.A. Bowers, and E. Edens (Eds.), *Improving Teamwork in Organizations* (pp. 265–282). Lawrence Erlbaum Associates, Mahwah New Jersey.

Flin, R., Martin, L., Goeters, K. M., Hörmann, H.J., Amalberti, R., Valot, C., & Nijhuis, H. (2005). Development of the NOTECHS (non-technical skills) system for assessing pilots' CRM skills (pp.133-154). In D. Harris and H.C. Muir (Eds.), *Contemporary issues in human factors and aviation safety*. Aldershot: Ashgate.

Flin, R., O'Connor, P., & Mearns, K. (2002). Crew resource management: Improving team work in high reliability industries. *Team Performance Management: An International Journal, 8,* 68-78.

Gaba, D.M., Howard, S.K., Fish, K.J., Smith, B.E., & Sowb, Y.A. (2001). Simulation-based training in anaesthesia crisis resource management (ACRM): a decade of experience, *Simulation and Gaming*, *32*, 175–193.

Gregorich, S.E., Helmreich, R. L., & Wilhelm, J.A. (1990). The structure of cockpit management attitudes. *Journal of Applied Psychology, 75* (6), 682-690.

Greif, S., & Kluge, A. (2004). Lernen in Organisationen [Learning in Organisations]. In H. Schuler (Ed..), *Enzyklopädie der Psychologie [Encyclopaedia Psychology], Themenbereich D, Serie 3, Wirtschafts-, Organisations- und Arbeitspsychologie [Business-, Organisational-, and Work-Psychology], Band 3, Organisationspsychologie - Grundlagen und Personalpsychologie* [Organisational Psychology – Basics and Personell Psychology] (pp. 751-825). Göttingen, Bern, Toronto, Seattle: Hogrefe.

Hagemann, V., & Kluge, A. (2014). Einflussfaktoren auf den Erfolg von und Methoden der Erfolgsmessung beruflicher Weiterbildung [Impact factors on the success of and performance measurement methods of professional training], *Wirtschaftspsychologie [Business Psychology], 16*, 81-93.

Hagemann, V., & Kluge, A. (2013). The Effects of a Scientifically Based Team Resource Management Intervention for Fire Service Teams, *International Journal of Human Factors and Ergonomics*, *2*, 196-220.

Hagemann, V. (2011). *Trainingsentwicklung für High Responsibility Teams [Training development for High Responsibility Teams]*. Lengerich: Pabst Science Publishers.

Hagemann, V., Kluge, A., & Ritzmann, S. (2011). High Responsibility Teams - Eine systematische Analyse von Teamarbeitskontexten für einen effektiven Kompetenzerwerb [A systematic analysis of teamwork contexts for effective competence acquisition]. *Psychologie des Alltagshandelns [Psychology of everyday activity]*, *4*, 22-42.

Helmreich, R.L., Merritt, A.C., & Wilhelm, J.A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology, 9*, 19-32.

Helmreich, R.L., & Wilhelm, J.A. (1991). Outcomes of crew resource management training. *The International Journal of Aviation Psychology*, *1*, 287-300.

Hollenbeck, J., Beersma, B., & Schouten, M. (2012). Beyond Team Types and Taxonomies: A Dimensional Scaling Conceptualization for Team Description. *Academy of Managment Review, 37*, 82-106.

Jackson, S.E., & Joshi, A. (2011). Work Team Diversity. In S. Zehdeck (Ed.), *APA Handbook of Industrial and Organizational Psychology Building and Developing the Organization* (Vol. 1, pp. 651-686). Washington: American Psychological Association.

Kluge, A., Sauer, J., Schüler, K., & Burkolter, D. (2009). Designing training for process control simulators: a review of empirical findings and current practices, *Theoretical Issues in Ergonomics Science, 10*, 489-509.

Kohn, L, Corrigan, J., & Donaldson, M. (2007). *To Err is Human. Building a safer health system*. Washington: National Academy Press.

Lauber, J. K. (1984). Resource management in the cockpit. *Air Line Pilot, 53*, 20-23.

McDonald, L.S., & Shadow, L. (2003). Precursor for Error. *An Analysis of Wildland Fire Crew Leaders' Attitudes about Organizational Culture and Safety*.

Third International Wildland Fire Conference/AFAC Conference Sydney, New South Wales, Australia.

Müller, M.P., Hänsel, M., Stehr, S.N., Fichtner, A., Weber, S., Hardt, F., Bergmann, B., & Koch, Th. (2007). Six steps from head to hand: A simulator based transfer oriented psychological training to improve patient safety. *Resuscitation, 73*, 137-143.

O'Connor, P., Jones, D. W., McCauley, M., & Buttrey, S.E. (2012). An evaluation of the effectiveness of the crew resource management programme in naval aviation, *International Journal of Human Factors and Ergonomics, 1*, 21-40.

O'Connor, P. (2011). Assessing the Effectiveness of Bridge Resource Management Training, *The International Journal of Aviation Psychology*, 21, 357-374.

O'Connor, P., Flin, R., Fletcher, G., & Hemsley, P. (2003). *Methods used to evaluate the effectiveness of flightcrew CRM Training in the UK aviation industry* (CAA Paper 2002/05). UK: Civil Aviation Authority (CAA).

Okray, R., & Lubnau, T. (2004). *Crew Resource Management for the Fire Service.* Tulsa, Oklahoma, U.S.A: PennWell Corporation.

Phillips, P., & Phillips, J. (2001). Symposium on the evaluation of training. *International Journal of Training and Development, 5*, 240-247.

Reason, J. (2008). *The Human Contribution*. Farnham: Ashgate.

Ritzmann, S., Hagemann, V., & Kluge, A. (2014). The training evaluation inventory (TEI) - Evaluation of training design and measurement of training outcomes for predicting training success, *Vocations and Learning, 7, 41-73*.

Ritzmann, S., Kluge, A., Hagemann, V., & Tanner, M. (2011). Integrating Safety and Crew Resource Management (CRM) Aspects in the Recurrent Training of Cabin Crew Members, *Aviation Psychology and Applied Human Factors, 1*, 45-51.

Röttger, S., Vetter, S., & Kowalski, J. (2013). Ship Management Attitudes and their Relation to Behavior and Performance, *Human Factors, 55*, 659-671.

Salas, E., Wilson, K.A., Burke, C.S., & Wightman, D. (2006a). Does crew resource management work? An update, an extension, and some critical needs. *Human Factors, 48*, 392-412.

Salas, E., Wilson, K., Burke, S., Wightman, D., & Howse, W. (2006b). Crew resource management training research, practice, and lessons learned. In R.C. Williges (Ed.), *Reviews of Human Factors and Ergonomics* (pp. 35–73). Thousand Oaks: Human Factors and Ergonomics Society.

Salas, E., Wilson, K.A., Priest, H.A., & Guthrie, J.W. (2006c). Design, delivery, and evaluation of training systems. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (pp. 472–512). Hoboken: J. Wiley.

Salas, E., Burke, C.S., Bowers, C.A., & Wilson, K.A. (2001). Team Training in the Skies: Does Crew Resource Management Training Work?, *Human Factors, 43*, 641-674.

Salas, E., Prince, C., Bowers, C.A., Stout, R.J., Oser, R.L., & Cannon-Bowers, J.A. (1999). A Methodology for Enhancing Crew Resource Management Training. *Human Factors, 41*, 161-172.

Schmutz, J., & Manser, T. (2013). Do team processes really have an effect on clinical performance? A systematic literature review, *British Journal of Anaesthesia, 110*, 529-544.

Sexton, J.B., & Klinect, J.R. (2001). The Link between Safety Attitudes and Observed Performance in Flight Operations. In *Proceedings of the Eleventh International Symposium on Aviation Psychology* (pp. 7-13). Ohio State University.

Sexton, J.B., Helmreich, R.L., Glenn, D., Wilhelm, J.A., & Merritt, A.C. (2000). *Operating room management attitudes questionnaire (ORMAQ)* (Technical Report No. 2). Retrieved from http://homepage.psy.utexas.edu/homepage/group/HelmreichLAB/Publications/595.doc

Thayer, P.W., & Teachout, M.S. (1995). *A Climate for Transfer Model* (AL/HR-TP-1995-0035). Texas: Brooks Air Force Base.

Weick, K.E., & Sutcliffe, K.M. (2003). *Managing the unexpected.* Stuttgart: Klett-Cotta.

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2004). Surgeons' attitudes to teamwork and safety. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 2045-2049). Thousand Oaks: SAGE Publications.

Can we remove the human factor from usability research to save time and money?

Andreas Espinoza & Johan Gretland
Sony Mobile Communications & Lund University
Sweden

Abstract

In today's corporate climate, managers look for different ways to cut corners. In such an attempt, this current research empirically evaluates the impact of taking the human factor out of usability research. The current study looks at whether expert users and functional performance (simple reduction of time and steps) can be of equal benefit to the usability refinement of a system compared to analysing real (novice) user performance. Four use cases are examined in the area of Near Field Communication (NFC) device connections. For the novice performance, 48 users attempted the 4 different use cases. Completion time, completion steps, user satisfaction ratings and user difficulty ratings are measured. The functional testing was an activity where system performance was objectively measured along with the performance of the optimal routes for each use case. The results indicate that a simple reduction in functional time and steps does not benefit the usability of the system and may actually be detrimental. While satisfaction and difficulty ratings correlate inversely with fewer steps and time, this primarily points to areas of necessary system design improvements indicated by human factors.

Introduction

What is the job of a usability tester? Our job is to impersonate a real user as much as possible - and when that is not enough – employ real users. For a current design, the optimal route means the non-improved best and fastest route to task completion. The user may choose the optimal route but that may not be fast enough or good enough e.g. when compared to competitor performance or a pre set criterion. Or users may fail to choose the optimal route, and thereby adding for instance time and steps. This would in turn indicate areas of possible usability improvements. Users may even be satisfied with the current usability performance (e.g. time and steps) thanks to the balance of system complexity and efficiency. This however can be controversial to management, especially when looking at and comparing with competitor products.

When management looks into usability, they learn that usability defines and comprises many different quantifiable quality traits, such as user satisfaction, ease of use and design efficiency. Clicks vs. time on tasks for instance, have some

In D. de Waard, J. Sauer, S. Röttger, A. Kluge, D. Manzey, C. Weikert, A. Toffetti, R. Wiczorek, K. Brookhuis, and H. Hoonhout (Eds.) (2015). Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference. ISSN 2333-4959 (online). Available from http://hfes-europe.org

correlation (Sauro, 2011), and reducing either of these can improve usability (ISO 9241-11, 1998). But all clicks are not made equal. Some clicks can be simple scrolling, while others are proper selections. Arbitrarily reducing the number of steps or time on task, does not necessarily improve the user interface – even if it is so stated in some literature (Zeldman, 2001) and by some "quant" usability people (Kieras, 2001).

But proper usability work often causes havoc with release dates and budgets. And it is difficult to convince others in the company that 5 users are enough, even if Jacob Nielsen (2000) says it is so, even if it saves time and money. The number 5 carries poor clout - or political force. The impression is that no truth can come from so few users. The discussion tends to slide towards methodology concerns rather than usability findings.

So why not let quality assurance (QA) people conduct the usability work? They are used to functional testing and can therefore easily count time and steps for the optimal route of a task, benchmark against a competitor or a set criterion, and propose necessary reductions in time and steps - to improve the usability. Management will say that a bug is a bug whether it is in the code or in the user interface design. But the problem is that it is unknown what a user bug is until a usability tester or an actual user tests the system. How can an existing bug be fixed when it has not been discovered yet? Only when applying the human factor, can these usability "bugs" be discovered, and the human factor is "missing" in QA testers – who are familiar with the system, and often are involved in its design. While quality certainly affects many usability traits (Nielsen, 2013), proposing that usability can be done by QA – or be combined with QA – is condescending to the usability field – and shows ignorance of what usability - and Human Factors – is.

The human factor is the magical dust that only a naïve user can provide, and which the usability professionals are willing to pay good money for. An entire team of thoroughly experienced usability researchers can spend hours analyzing a new system and exposing many usability problems, but they can fail to expose a critical finding that e.g. one 16 year old girl will point out in 10 minutes. If it is considered controversial to reuse test participants for several tests since they become familiar with the system and the setting, it must be a cardinal sin to employ QA people for usability testing.

Curiosity, or perhaps self preservation, enticed the examination of this area further. The goal is to understand if there can be enough value in usability by functional test to simply remove the user all together. The answer is no of course, but the practice of desktop usability conducted by untrained personnel is widespread and growing. It is cheap, fast and basically anyone can do it with very little direction. Reliability can be very strong adding to the problem, while validity is nonexistent. If you repeat the same non truth long enough and loud enough, someone will believe you. This is especially problematic when the concepts of scientific method and validity are left "on the cutting room floor", so to speak.

One-Touch is the SONY feature name for the process of connecting two devices using NFC by touching them together for a short time (NFC Forum, 2014). This

study examined the functional performance for this connection procedure, as well as the user performance. The purpose of this study was to compare functional performance of an expert user using the fastest route (Expert Performance, EP), with that of a novice user unaware of the optimal route (Novice Performance, NP). The discrepancies in expert and novice performance are attributed to usability findings (human factors) which can be converted into actionable requirements. These requirements can then be implemented by development teams to improve the user performance and user experience. The four use cases in table 1 were examined:

Table 1. The four One-Touch use cases examined

1)	One-Touch Mirroring
2)	One-Touch Music
3)	One-Touch Sharing
4)	One-Touch Connection

This report describes the testing procedure, the test results and a comparison of the usability issues (usability findings) between EP and NP. A presentation of the fastest routes for each use-case is included as well as the EP and NP values (time and steps). The report also presents a comparison of competitor benchmark measurements, target proposals (time and steps) for each use-case and actionable requirements tied to the usability findings. A discussion is presented around the topic of replacing user testing with functional testing. Furthermore, a task analysis was conducted to understand in detail the requirements necessary for the users to perform their tasks and achieve their intended goals, and it is presented in the discussion section.

Method

A traditional usability test was conducted using the SONY Lund Experience Lab. Four use cases were examined and the independent variables were the different phones used in each use case. The dependent variables were completion time, completion steps, user satisfaction ratings and user difficulty ratings. The functional testing (Expert Performance) was a desk activity where system performance was objectively measured.

Participants

For the EP measurements the authors themselves acted as experts and measured the performance of the optimal routes for each use-case. The authors computed the optimal route, minimal number of steps and measured shortest possible time to achieve the task.

For the NP measurements external participants were recruited. There were 48 external participants in total, of which 24 were men and 24 were women. The average age was 24. Most participants were university students with iPhones or other smart phones, unfamiliar with the One-touch concept. Tables 2, 3 and 4 describe users' characteristics.

Table 2. Participants divided into occupation

Occupation	
University student	43
Nurse	3
Doctor (MD)	1
Journalist	1

Table 3. The number of participants using a certain phone

Current Phone Used	
iPhone	26
Samsun Galaxy	8
Sony Xperia	6
Non smart phone	4
HTC One	3
Nokia Lumia	1

Table 4. Number of participants familiar or experienced with the One-touch concept

One-touch familiarity			
Used before		Heard about	
Yes	No	Yes	No
2	46	6	42

Material

The SONY Lund Experience lab was used. The equipment used in the test is listed in table 5. Due to confidentiality issues it is not possible to reveal specifics about the SONY prototype products used in the study.

Table 5. The equipment used during the usability test

1) SONY1 AP1 (prototype phone)
2) HTC One (sales model phone)
3) Samsung Galaxy S4 (sales model phone)
4) LG G2 (sales model phone)
5) SONY2 AP2.1 (prototype phone)
6) SONY Bravia 40w905a (prototype TV, final build)
7) SONY SRS-BTM8 NFC and Bluetooth speaker (sales model speaker)
8) SONY Smart Watch 2 PQ (prototype smart watch)

A wall mounted lab camera (unknown brand) was used to record the user while performing some tasks. In table 6 the equipment used for each use-case is described.

Table 6. A description of equipment used for each use-case

Use-case	One-Touch Mirroring	One-Touch Music	One-Touch Sharing	One-Touch Connection
Phones Used	Sony1 Sony2	Sony2 LGG2 HTC One Samsung S4	Sony2 LGG2 HTC One Samsung S4	Sony2 LGG2 HTC One Samsung S4
Receiving Equipment	Bravia TV	SRS-BTM8	Sony1	Sony Smart Watch

All devices were updated with the latest available software/firmware. The phones were loaded with identical music and video content to be used during testing. The Walkman application was used for SONY products and the default audio/video player for competitor products.

Procedure

This study was conducted in two parts.

The first part was the expert performance (EP) testing, where completion time and steps for each use-case was measured. The EP testing followed the optimal and most efficient route to task completion. These are the fastest most efficient routes, requiring the least amount of steps and the least amount of time to complete. Each phone was verified 5 times. The optimal routes are presented in figure 1. The number in the first box indicates the total number of steps.

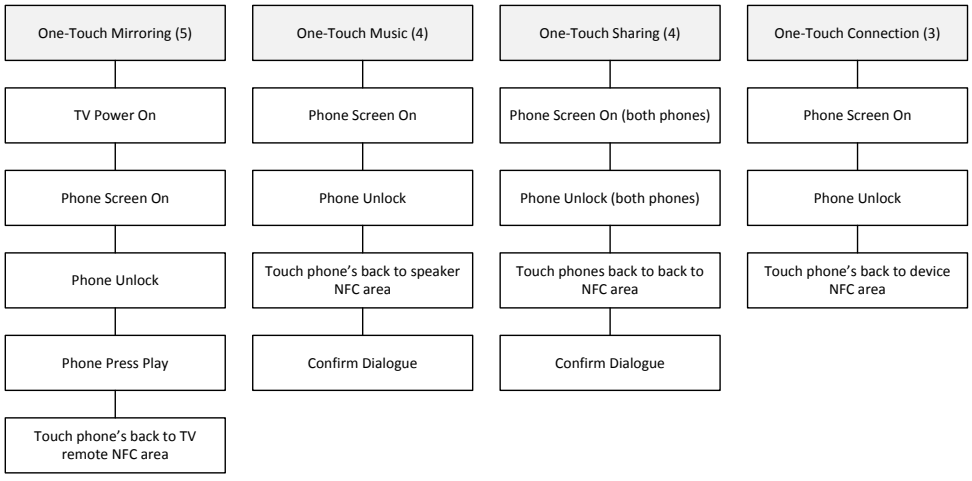


Figure 1. The optimal routes for each use case.

The second part was the novice performance (NP) testing.

All participants signed confidentiality agreements prior to the start of the test.

The NP testing had identical device preconditions (table 7) as the EP testing, but the users were novice and were presented with training material prior to test start (table 8). The training material used was the official SONY marketing videos for each feature.

Table 7. Test device preconditions

Mirroring	Film is ready but not playing. Phone on with screen off.
Music	Music is playing. Phone on with screen off.
Sharing	Image is ready. Same image used for all devices. Phone on with screen off.
Connect	Smart Connect app installed. Phone on with screen off. Smart Watch off.

Table 8. Training material for each use case presented to the user prior to testing

One-Touch mirroring	https://drive.google.com/folderview?id=0B13MwD-rhB5ZQ2trV1pJcDhkXzA&usp=sharing
One-Touch music	https://drive.google.com/folderview?id=0B13MwD-rhB5ZWEdwbHJuUnFTSWc&usp=sharing
One-Touch sharing	https://drive.google.com/folderview?id=0B13MwD-rhB5ZZzVCZzhKbmhESlU&usp=sharing
One-Touch connection	https://drive.google.com/folderview?id=0B13MwD-rhB5ZUU1DdVVoM0JjbUk&usp=sharing

The participants were divided into four groups of 12 in each group. Each group was assigned one use-case (table 1). All participants in each group attempted the assigned use-case for all available devices. To cancel any learning effects, a predetermined order using a Latin Square configuration was used (see table 9). For the use-case One-Touch mirroring, only some SONY Xperia devices were compatible with the SONY TV being tested. No competitor devices were compatible with the SONY TV.

Table 9. A Latin Square configuration was used to cancel any learning effects

Participant	First	Second	Third	Fourth	Order
1	SONY1	LG	HTC	Samsung	A
2	Samsung	SONY1	LG	HTC	B
3	HTC	Samsung	SONY1	LG	C
4	LG	HTC	Samsung	SONY1	D
5	SONY1	LG	HTC	Samsung	A
6	Samsung	SONY1	LG	HTC	B
7	HTC	Samsung	SONY1	LG	C
8	LG	HTC	Samsung	SONY1	D
9	SONY1	LG	HTC	Samsung	A
10	Samsung	SONY1	LG	HTC	B
11	HTC	Samsung	SONY1	LG	C
12	LG	HTC	Samsung	SONY1	D

There were 4 different EP and NP tests – one for each use-case. The test duration was one week for each use-case.

Quantitative data was collected for the EP and NP tests. The number of steps and time it took to complete each task was recorded. Subjective data ratings for difficulty and satisfaction were collected for the NP tests using the Single Ease Question (SEQ) test, a validated post test questionnaire (Sauro, 2010). Moreover, qualitative data was collected for the NP tests. The moderator observed where the users had problems and made note of these problems. The moderator also helped the users to express what they were thinking, by often asking them "what are you thinking now?"

Results

For expert performance, the following table 10 indicates the number of task steps and mean task time, where time is divided into the respective parts. One part is the time it takes to perform the necessary steps. The other part is the total time it takes to complete the action, including time due to system elaboration.

Table 10. Number of steps, mean step time (user) and mean total time (user+system) for expert performance (EP)

[steps], mean time for steps (s) / total time (s)				
One Touch	Mirroring	Music	Sharing	Connect
SONY1	[5] 11/17	[4] 6/29	[4] 6/15	[3] 5/13
Samsung	-	[4] 9/27	[4] 12/25	[3] 6/16
HTC One	-	[4] 9/28	[4] 7/22	[3] 5/13
LG G2	-	[4] 9/27	[4] 8/21	[3] 6/12
SONY2	[5] 11/20	[4] -/22	[4] -/16	-

Table 11 indicates the mean novice performance (steps and mean total time) for all phones and use cases.

Table 11. Number of steps and mean total time for novice performance (NP)

step/time(s)	One-Touch			
	Mirroring	Music	Sharing	Connect
SONY1	10/87	12/90	8/40	10/79
Galaxy S4	-	15/112	8/63	14/102
HTC One	-	10/75	14/86	11/90
LGG2	-	16/115	12/93	13/103
SONY2	12/99	-	-	-

Table 12 outlines failure rates for novice performance for all phones and use cases. Failure time limit was set at 240 seconds. No failure step limit was set.

Table 12. Novice performance (NP) failure rates

%	Failure Rates			
	Mirroring	Music	Sharing	Connect
SONY1	33	25	8	0
Galaxy S4	-	66	25	8
HTC One	-	58	17	42
LGG2	-	83	25	17
SONY2	0	-	-	-

Subjective data ratings for difficulty and satisfaction were collected for the NP tests using the SEQ test; a 7 point scale where 1 is very bad and 7 is very good. The results are plotted against time and steps respectively. Please see below figure 2 thru 5 for ratings for satisfaction and difficulty.

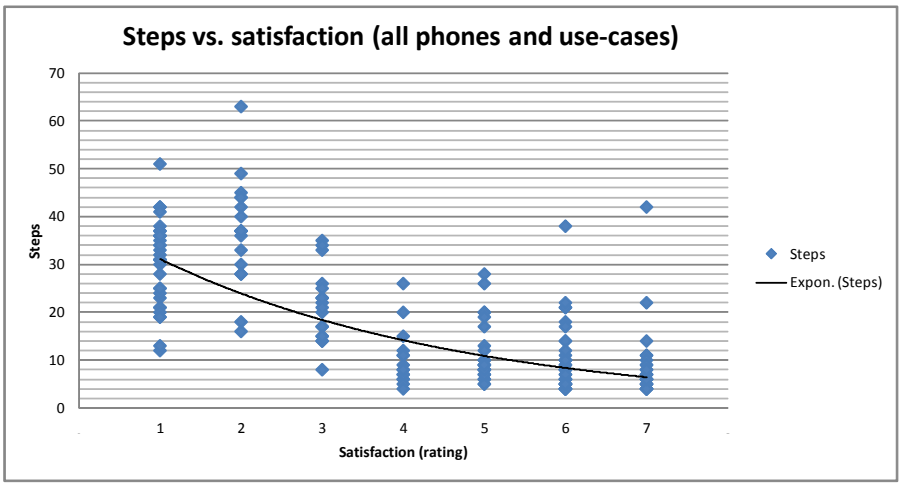


Figure 2. Subjective data for satisfaction compared with number of steps.

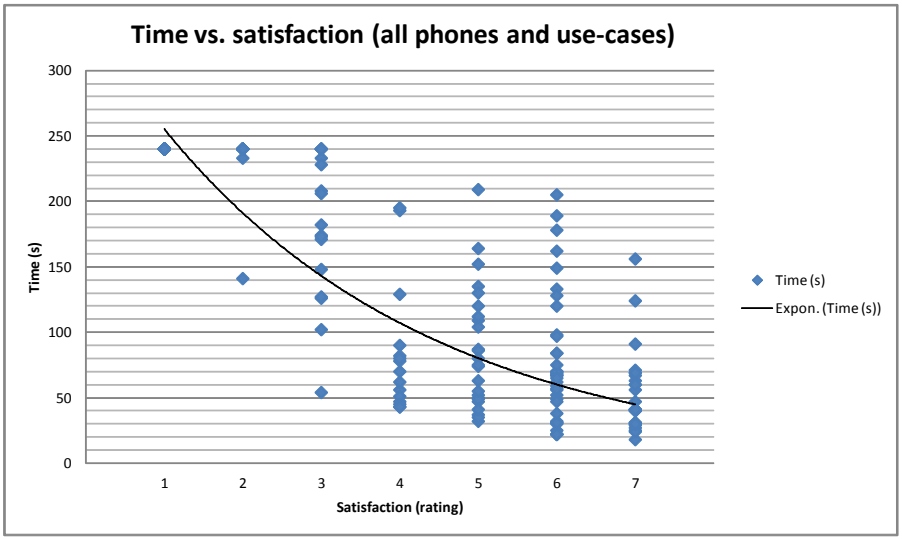


Figure 3. Subjective data for satisfaction compared with time.

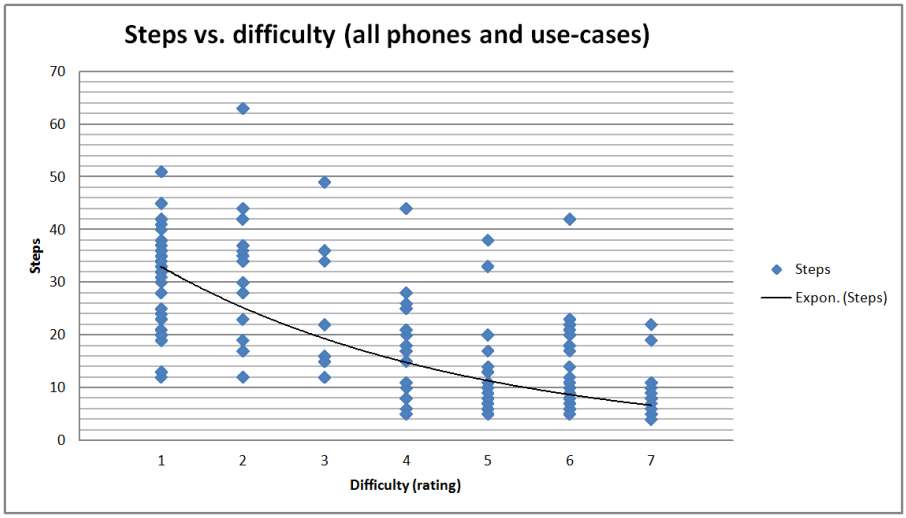


Figure 4. Subjective data for difficulty compared with steps.

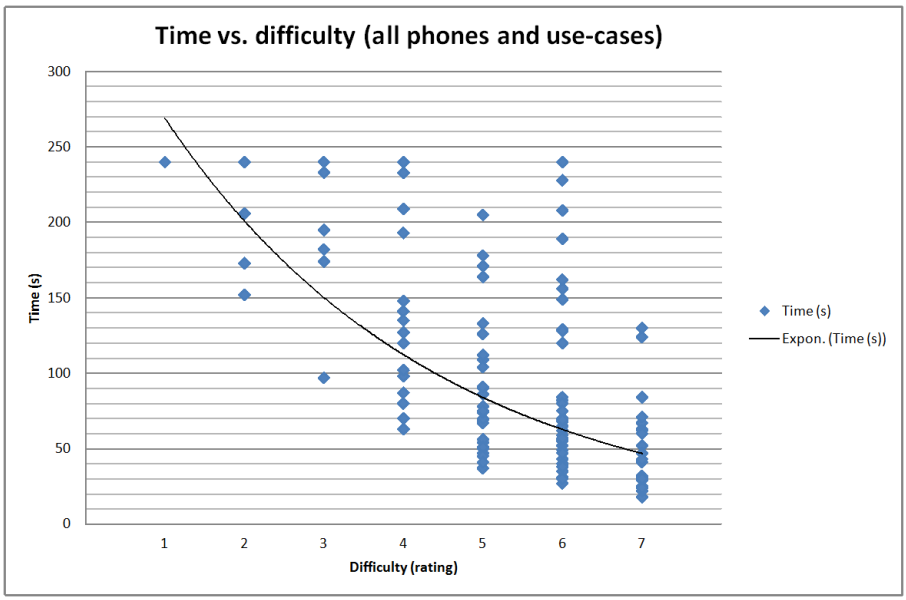


Figure 5. Subjective data for difficulty compared with time.

Usability findings

In table 13 the major usability findings and proposed actionable requirements are described. The usability findings are a key contributing factor for the NP deviations from the EP performance figures. Due to space limitations only an excerpt of the 16 main findings is presented.

Table 13. Excerpt of the 16 usability issues and actionable requirements for all use cases

Usability Findings (General for all Use cases)		**Actionable Requirements**	
UF1	Users don't understand that they need to physically touch the devices together.	AR1	SONY must clearly communicate the need to physically touch devices. A manual, wizard, animation or small instruction film should be included to demonstrate the functionality.
UF2	Users have problems aligning the NFC transceivers. Users don't know about NFC and they don't recognize the icons. They randomly touch the devices together. They don't understand that they have to touch them on an exact NFC location.	AR2	The NFC symbol should be present and clearly detectable on all NFC devices, indicating transceiver placement. A manual, wizard, animation or small instruction film should be included to demonstrate the feature.
UF3	Users don't understand that for NFC to be active the screen needs to be lit and the phone unlocked.	AR3	Transfer should work all of the time, or, transfer should work at least for a few seconds after the screen has locked or is automatically turned off.
UF16		AR16	

EP vs. NP differences in human factors

Table 14 illustrates the performance difference when comparing EP with NP by presenting the mean for all use cases and users. Included are also the number of usability findings as discovered through functional testing and user testing.

Table 14. Performance and usability issue difference between EP and NP

	EP from functional testing	NP from user testing
steps (average - all cases)	4	12
time (s) (average - all cases)	17	89
# of concrete actionable usability findings (total all cases)	0	16

Performance targets

Performance targets are proposed in table 15. Implementation of technical enhancements as well as interaction design improvements based on usability findings in this test will help ensure that the system delivers a satisfactory and competitive user experience. It is important to understand that these values indicate the time and steps for the entire use case, which includes functional system time and user interaction, for first time usage. The targets are based on the following acceptance criteria:

1) Achieving at least an average user rating of 6 on the 7 point SEQ scale as illustrated in figures 2 thru 5 above

2) Matching values equivalent to the best competitor product
3) Exceeding current value with 10% when SONY is the best performer for the use-case
The following UX targets are proposed:

Table 15. Proposed performance targets based on acceptance criteria

	One-Touch:			
TARGETS	**Mirroring**	**Music**	**Sharing**	**Connect**
Average time	70 s	70 s	36 s	67s
Average steps	8 steps	10 steps	7 steps	9 steps

Task Analysis

The following task analysis (figure 6) presents all the necessary information and knowledge the novice user needs in order to succeed at expert performance levels. It includes general knowledge applicable to all use cases and specific knowledge applicable to specific use cases (in parentheses). In comparison, the expert performance levels solve the tasks with the optimal route and perform with the least amount of time and steps. The expert performance was attained in this study when applying optimal routes for each use case, as illustrated in figure 1.

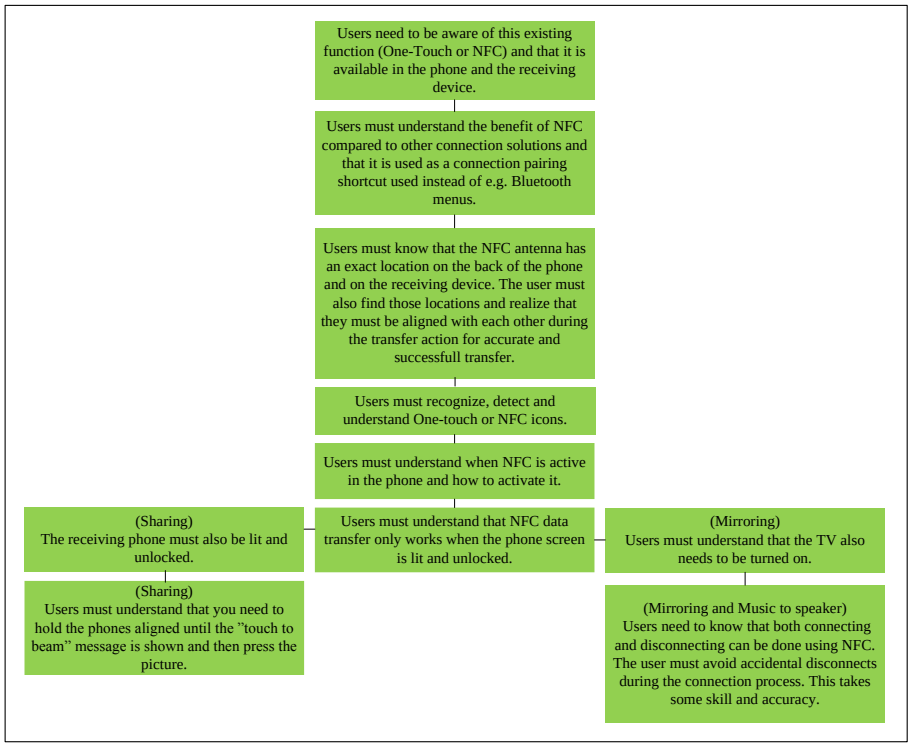


Figure 6. Task analysis for all uses-cases, showing general and specific knowledge requirements.

Discussion

When considering comparisons between functional testing and user testing, it is not really important what the performance differences are between the phones – i.e. in the benchmark task. What is important to pay attention to, is that there is a large discrepancy between expert performance and novice performance. The cause of this discrepancy is attributed to human factors. Task failure, as noted in table 12, does not exist for the functional testing, which in itself is a serious indication of the problems with using functional testing instead of usability testing.

The results from the SEQ surveys are much as expected. A lot of users are unhappy – but interestingly a lot of users are happy too. It’s a peculiar management concern how time and steps affect user satisfaction and difficulty, when users actually have higher tolerance for time and steps than current optimal expert performance numbers, which management still want to decrease. The reason is the concern with competitor performance. One can see the same thing for both time vs. satisfaction and steps and time vs. difficulty. Management focus is on a reduction of time and steps to be competitive in a benchmark situation when a lot of users are nowhere near the expected expert performance results and a large group is well satisfied with far lower performance (see e.g. figure 5).

Functional testing using for instance QA (quality assurance) people not only fails in understanding what the user needs to succeed but also what level of performance users are satisfied with. This translates into pouring many wasted dollars into creating a stronger performing system when users are satisfied with a lesser performing system where money needs to go into solving human factors problems - allowing more people to succeed instead.

References

ISO, (1998). ISO 9241-11:1998 Ergonomics of Human System Interaction. Geneva, Switzerland: ISO.

Kieras, D. (2001). Using the keystroke-level model to estimate execution times. Retrieved from http://courses.wccnet.edu/~jwithrow/docs/klm.pdf

NFC Forum (2014). Connection handover user experience recommendations. [White paper] Retrieved from http://nfc-forum.org/wp-contetnt/uploads/2014/09/ConnectionHandoverUserExperience-White-Paper-Sept14F1.pdf

Nielsen, J. (2000). Why you only need to test with 5 users. Retrieved from http://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

Nielsen, J. (2013). QA & UX. Retrieved from http://www.nngroup.com/articles/quality-assurance-ux/

Sauro, J. (2010). If you could only ask one question, use this one. Retrieved from http://www.measuringusability.com/blog/single.question.php

Sauro, J. (2011). Click versus clock: Measuring website efficiency. Retrieved from http://www.measuringusability.com/blog/click-clock.php

Zeldman, J. (2001). Hierarchy and the so called 3-click rule. In K. Whitehouse and S. Kearns (Eds.), *Taking your talent to the web - a guide for the transitioning designer*. (pp 97-99). San Francisco: New Riders Publishing.

Influence of head mounted display hardware on performance and strain

Matthias Wille & Sascha Wischniewski
German Federal Institute for Occupational Safety and Health (BAuA)
Germany

Abstract

In high reliable industries, where critical information is often given in real time, Head-Mounted Displays (HMDs) may support workers. Especially if mobility is needed, like in maintenance or some fields of medicine, HMDs can display this critical information directly within the field of view. However, in a study presented on last year´s conference of HFES-Europe we showed that participants react less accurate to a monitoring task presented on an HMD compared to a Tablet-PC, although the information was displayed always within their sight. These results might be based partially on performance decrements caused by the additional strain from handling the uncomfortable and heavy industrial HMD. In a new study we replicated the experiment with the new, lighter and more comfortable consumer HMD Google Glass to investigate the influence of hardware on performance and strain. Results show some significant improvements in HMD technology regarding reported comfort: some visual fatigue items were rated lower and less headache and neck pain were caused by the HMD. But the performance in an assembling task and parallel monitoring task still is worse on HMD compared to Tablet-PC. This implicates that displaying critical information on an HMD might not help to draw the user's attention.

Introduction

Head-Mounted Displays (HMDs) have become more affordable and comfortable during the last years and are now on the cusp of mass-market. Beneath applications in the consumer world HMDs may also support workers: work relevant information can be shown within the field of view while both hands are still free for a manual task. Possible applications can be found for instance in maintenance, assembling, logistics, some fields of medicine (e. g. anaesthesia, where the patient and some relevant data have to be monitored at the same time) or in police and rescue teams. Whenever information is needed during a work process and mobility and hands free are also an axiomatic features HMDs can be a solution to support workers. But there are still many questions remaining if it comes to prolonged work with HMDs and therefore the German Federal Institute for Occupational Safety and Health (Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, BAuA) started a research project focussing on different aspects of work assistance by HMDs.

Project

The project, where both of the studies mentioned in this paper took place, is titled "head mounted displays – conditions of safe and strain-optimised use" and has different work packages with diverse goals. One work package focusses on task analysis and the question in which situations an HMD might be appropriate (Grauel, Adolph, & Kluge, 2014). Here one main result states that the interdependency between individual worker, technology and task is crucial and therefore common statements are hard to give. Another work package focusses on the physical strain during prolonged work with an HMD (Theis, Alexander, Mertens, Wille, & Schlick, 2014; Theis, Alexander, Mayer, & Wille, 2013): in a 3.5 hour session subjects were deconstructing and constructing a real car engine while information was shown on an HMD (Liteye 750 A) or on a wall mounted monitor. Main results showed by comparing pre - and post - tests no influence on the visual system (visual acuity, peripheral field of view, eye blink - rate and - duration measured by EOG). Within muscle activity of neck and shoulder – measured by EMG during the whole session – only the left M. Splenius capitis showed a higher increase over time with HMD and in a video analysis less head movements while working with HMD was proven. In questionnaires about visual fatigue a higher increase over time for HMDs in values like "heavy eyes", "neck pain" or "headache" was found. Although there is some higher strain in muscle activity there is no physical no-go-factor for using HMDs during prolonged work. But results also show 10-18% higher work execution times with the HMD compared to the wall mounted monitor. In another work package the two studies reported here were conducted with focus on mental strain and performance. Subjects had to fulfil a graphical assembling task and parallel react to a monitoring task. Main results showed higher subjective strain ratings with HMD compared to a Tablet-PC and also higher work execution times. However, it is worth mentioning that after a phase of habituation objective strain parameters (heart rate and heart rate variability) showed no differences between display types (these results are not published yet). An overview on the project is also given in Wille et al., (2014). In the end of the project implications for occupational safety and health while working with HMDs will be carved out and also hints for the risk assessment of work places using HMDs will be given. More information about the project and a complete list of publications can be found here: http://www.baua.de/en/Research/Research-Project/f2288.html [January 2015].

Scope of this paper

In this paper we would like to compare two studies that used the same task but different kind of HMDs. As technology evolves rapidly it is important to investigate how much of the effects are based on the technology itself (the fact of having a monocular near to eye display) or rather on the current available hardware implementation. Hardware bought at the beginning of a research project can be already antiquated when results are published. In case of HMD it is possible that some amount of the strain is more based on e. g. heavy head carriers than on working with an HMD in common. A comparison of both studies will show.

Method

In this section an overview of the method of both studies that are compared here is given. For further details please refer to the former publications (industrial HMD study: Wille, Grauel, & Adolph, 2013; consumer HMD study: Wille, Scholl, Wischniewski, & Van Laerhoven, 2014).

Experimental Design

Although both studies use the same tasks (as described below), collect parameters at same timestamps and to some amount also use the same participants, there are some differences in the design that should be mentioned in the beginning for better understanding (Figure 19):

The industrial HMD study (which was the first study) used a within subject design: each subject came three times and worked for 4 hours each. Two times they worked with the MAVUS-HMD (second HMD trial was done for investigating habituation to the technology) and one time with a Tablet-PC (comparison with second HMD trial to investigate influence of display technology after an eventual habituation to the HMD).

In the consumer HMD study a between subject design was used. So there was only one session and half of the subjects worked with Google Glass while the others worked with a Tablet-PC. Furthermore, this replication study was much shorter (only 30 minutes) and embedded in a series of studies conducted together with the chair of embedded sensing systems (EES) of the University Darmstadt. Participants worked about 1 ½ hour in other studies and on other tasks before this replication was done in the end of the about 2 hours long study series. A complete replication where participants work for 4 hours again was due to organisational aspects not possible.

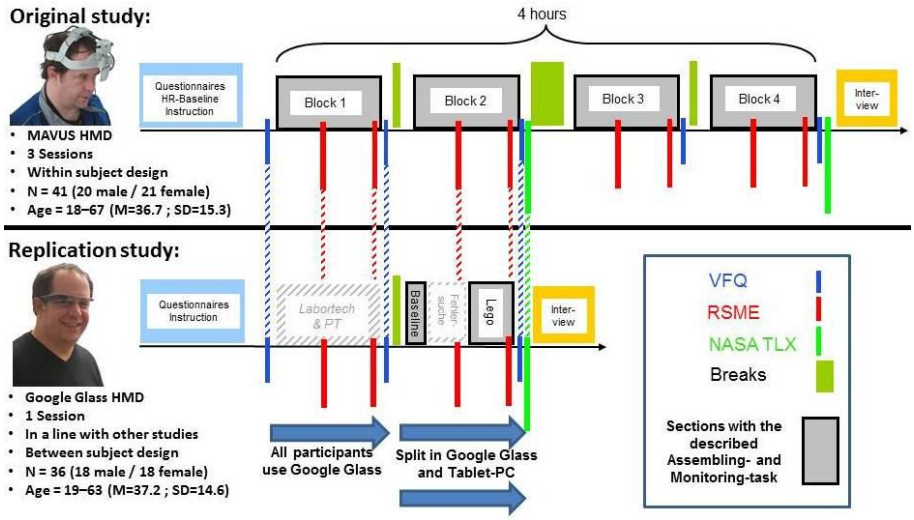


Figure 19. Experimental design of both studies. For details see text.

Participants

41 participants took place in the industrial HMD study. They were aged between 18-67 (M = 36.68; SD = 15.285; 20 male and 21 female). None of the subjects had worked previously with an HMD.

In the consumer HMD study 36 subjects participated aged from 19-63 (M = 37.16; SD = 14.643; 18 male and 18 female). 30 of these subjects had also participated in the industrial HMD study about 8 months before and therefore had some minor knowledge of HMDs and the tasks.

16 participants were selected for the "direct comparison of subjective strain on both HMDs" presented at the end of the paper. All these subjects took part in the industrial HMD study and they were in the Google Glass group of the consumer HMD study. They were 20-63 years old (M = 38.88; SD = 14.64) and 6 male and 10 female. The unbalanced gender should be no problem as no gender effects showed up in both studies.

Tasks

In both studies the same task combination had to be done: a dual task paradigm, weighted by instruction as equally important and both to be handled as fast and accurate as possible. Subjects had to build up toy cars, given a graphical step by step instruction based on Lego-Technic, and parallel supervising a monitoring task that was presented on the peripheral border of the screen (Figure 20).

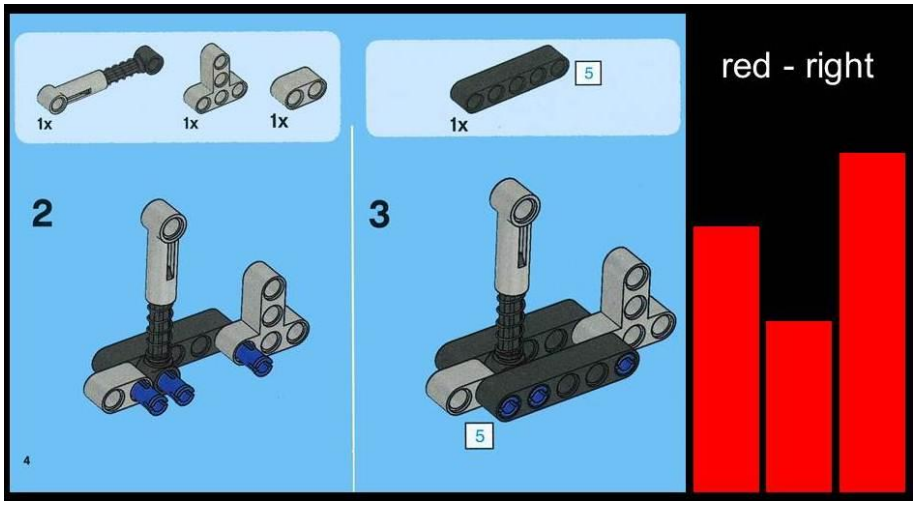


Figure 20. Work content as presented on the displays. Here from the Google Glass study (as shown in Wille et al. 2014). On the left side the assembling task based on Lego-Technic. On the right side the monitoring task with 3 bars and feedback about last confirmed color and position on top. Monitoring task was always presented on the exterior side of the display. On the industrial HMD the background was grey and the format was 4:3. The Assembling slides were fitted to the screen format.

The assembling task was selected because its sequential, graphical character is in line with some real life working tasks for instance in maintenance. The former Lego-Technic slides were all fitted to the screen size and some of them were rearranged or expended with arrows for better visual ergonomics. As a dependent variable the number of processed slides was used because the slides are comparable regarding work demand and more complex models just contain more slides. Errors in assembling were not tolerated and mostly noticed by the participants themselves as bricks did not fit in later slides and therefore subjects have to step back. In the industrial HMD study participants built up many different models and every time they finished one immediately the next one was given, to get about four hours of continued work. In the consumer HMD study, where only 30 minutes of time were intended, subjects only had one model which was not finished in this short time. Here the number of processed slides within 25 minutes was the dependent variable for performance.

The parallel presented monitoring task consists of two tasks within: on one hand the three bars changed their colour (red-blue) from time to time and subjects were asked to confirm this. This type of monitoring task has a visual pop-out effect as some amount of the screen changes colour at once. On the other hand the bars changed slowly but continuously and independent from each other their length and subjects had to confirm each time the position of the longest bar changed within the three bars. This monitoring task is harder to detect as it got no visual pop-out. All these variations were random based. As seen in a later analysis of data in the industrial HMD study the colour change happened about every 140 seconds and the change in position of the longest bar every 95 seconds. While in the consumer HMD study the colour change happened about every 106 seconds and length change every 94 seconds.

Apparatus and interaction

In the first study HMD and Tablet-PC were industrial products. Those products which are mainly used in industrial environment are more robust and have higher tolerances regarding humidity and temperature. Furthermore the accumulator mostly holds longer than on consumer products. The MAVUS-HMD from the Heitec company is a monocular look-around display with a resolution of 800 x 600 pixels (figure 3, left). The technique is fixed to a head carrier which includes a front camera and a headset. But camera and headset had no function during the study, while in industrial applications they are provided for communication. The head carrier including all technology weighted 380 grams and was cable connected with a vest including the radio technology for the transmission of data and the accumulator for power supply. As Tablet-PC the CL900 by Motion was used with a screen size of 10`` and weight of 950 grams. To ensure that representation of the work content was comparable, only a window of 800 x 600 pixels was shown and the rest of the area was covered. In this study all interactions on both devices - switching the construction slides forward and backward and confirming the monitoring tasks - took place via a converted number pad.

In the Google Glass replication study HMD and Tablet-PC were consumer devices. These are mostly a bit lighter and fancy but also less robust and have less tolerance

concerning humidity and temperature. Furthermore the accumulator often has less power: On Google Glass for instance, which is designed for micro interactions of only a few seconds, the battery will keep up less than one hour if continuous information is displayed (like in our setup). However, those consumer products may give an idea of how products for work situations might look very soon. Google Glass is a 640 x 360 pixel see-through display mounted on a spectacle frame weighting 50 grams (figure 3, right). It was connected to a battery extension pack to enable continuous displaying of information for about 2 hours. The Tablet-PC was a Samsung Galaxy SM-T210 with a resolution of 1024 x 600, a screen size of 17.8 cm (7'') and a weight of 300 grams. In this second study all interactions on Google Glass were done by speech commands and all interactions on the Tablet-PC were done by touch. Speech commands were: "next slide" and "previous" for changing the assembling slides and "bar changed" for both monitoring tasks. On Google Glass an additional zoom function was given by saying "zoom image" that enlarge the assembling image twice while the presented part was chosen with head movement measured by internal sensors. To shrink the image the speech command "scale down" was used. On the Tablet-PC a swipe to the left opened next slide and a swipe to the right the previous slide. A double tap (anywhere on the screen) was used as confirmation in the monitoring tasks. On the Tablet-PC no zoom function was given.

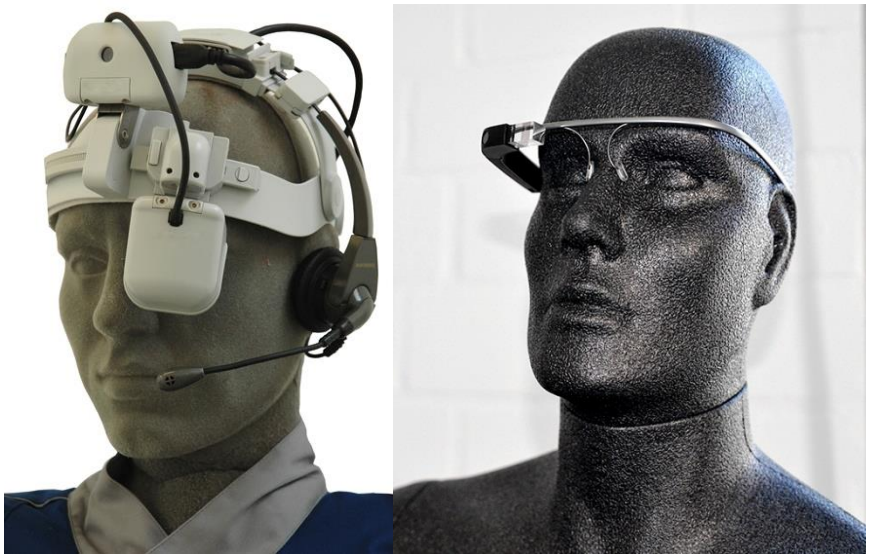


Figure 21. The used Head-Mounted Displays. Left: MAVUS-HMD. Right: Google Glass.

Dependent Variables

In both studies some dependent variables were collected at same timestamps (figure 1). The NASA-TLX (Hart & Staveland, 1988) as a well-known subjective strain questionnaire after 2 hours (and after 4 hours in the industrial HMD study). The Rating Scale of Mental Effort (RSME, Zijlstra, 1993) was collected every 30 minutes in both studies. Subjects were introduced in this scale before the experiment

started and collection was done during the work process. The Visual Fatigue Questionnaire (VFQ, Bangor, 2000) is a questionnaire with 16 items about visual fatigue like “irritated/burning eyes”, “difficulties to see sharp”, but also asks for headache, neck pain and mental fatigue and let subjects rate all items on a 10 point scale. This questionnaire was collected before the beginning (to investigate the individual initial position) and after each hour at the beginning of the breaks. The performance in the assembling task was characterised by the number of processed slides within the experiment time. In the monitoring tasks the hit rate (percentage of appropriate reactions to stimuli) and the reaction time of appropriate reactions were dependent variables.

Results

In this section the main results of both studies are presented and compared. For a more detailed analysis please refer to the original papers or the upcoming project report (German language). Furthermore some of the effects slightly differ in amount from the original papers. For the consumer HMD study this is based on the full sample (N=36) now, while the paper had only 20 subjects in the on-going study at that time. For the industrial HMD study complexity was reduced for this comparison: Only performance of the second HMD session is reported here (to counteract possible habituation during first HMD session) and in the monitoring task only trials with given feedback were reported. In original study feedback was a factor and only given in half of the trials. Age of subjects as factor was also cut here as the sample size for the direct comparison (N=16) is too small for another factor and same subjects are analysed with same age distribution.

Construction task

Both studies show that with the HMD less assembling slides were done in same time compared to the Tablet-PC. In the industrial HMD study during the 4 hours in mean 129 slides were done with HMD and 158 slides with Tablet-PC. This significant effect [$F(1, 40) = 25.944$, $p < .001$] means that 22.5 % more slides were conducted with the Tablet-PC. It is worth noting that for the comparison the second trial with HMD was used, where subjects had before another 4 hour session with HMD to get some habituation. In the consumer HMD study during the 25 minutes in mean 17.1 slides were done with the HMD and 22.9 slides with the Tablet-PC. This also significant effect [$F(1, 35) = 5.725$, $p = .022$] means that 33.9 % more slides were conducted with the Tablet-PC. Both findings are in line with the longer task execution time on HMD found by Theis et al. in the study about physiological strain with HMDs.

Monitoring tasks

In the industrial HMD study no significant difference could be found between display types [$F(1, 40) = 2.583$, $p = .116$] but hit rate on the Tablet-PC was better than on HMD (figure 4). Furthermore there is a significant effect of task type [$F(1, 40) = 89.897$, $p < .001$] indicating better reaction to the colour change task (which was expectable based on the visual pop-out effect).

For the consumer HMD study it has to be stated that the hit rate on colour change is not 100% trust worthy: A hidden process in Android wrote erroneous colour change events into the results matrix which resulted in an increased number of misses even though the participant was not able to react. The problem occurred on HMD and Tablet-PC in the same way and may increase the misses about 10%. This might also explain why here no effect of task type can be found [$F(1, 35) = .108$, $p = .744$]. The effect of display however becomes significant this time [$F(1, 35) = 5.337$, $p = .027$] with clearly worse hit rates on the HMD. Furthermore, in this study a baseline was carried out where the monitoring task was conducted as single task (without parallel assembling) for 5 minutes at the beginning of the second hour (Figure 19). Here a clear significant effect [$F(1, 35) = 18.249$, $p < .001$] was found, proving that reaction was more accurate during single task than during dual task, which was expectable too.

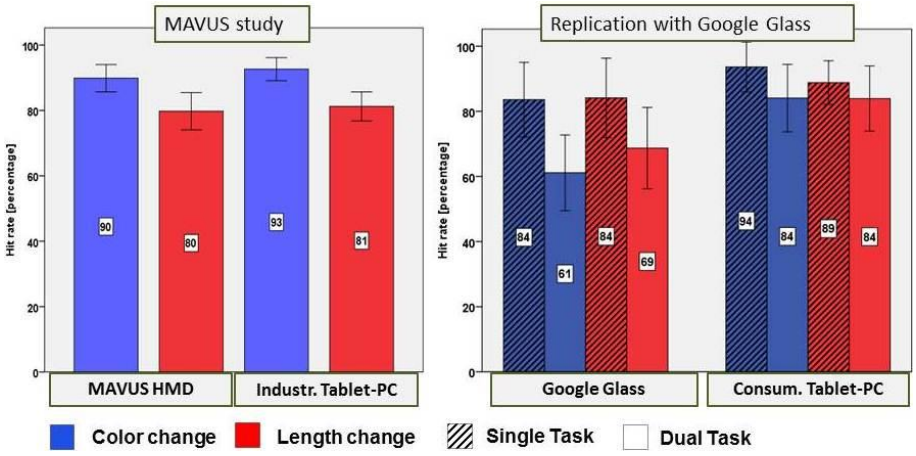


Figure 22. Hit rate in both monitoring tasks (colour change and length change) for both studies. Whiskers represents the 95% Interval.

Table 14. Average reaction time in seconds for the monitoring tasks by display (standard deviation in parentheses)

Reaction times	*MAVUS-HMD*	*Industrial Tablet-PC*	*Google Glass*	*Consumer Tablet-PC*
Colour changes	12.9 (8.98)	11.5 (6.65)	9.4 (4.84)	5.5 (3.56)
(as single task)	-	-	3.8 (4.26)	1.7 (0.82)
Length changes	20.0 (14.18)	16.1 (10.62)	9.7 (5.19)	8.6 (10.15)
(as single task)	-	-	3.1 (3.01)	3.8 (3.83)

Table 14 shows the reaction times for correct reactions in monitoring tasks for all displays. If comparing one should keep in mind that the type of interaction was not the same under all conditions and this might have an influence on reaction time too: During MAVUS-HMD and industrial Tablet-PC sessions all reactions were done manually on a converted number pad by pressing buttons. Furthermore here the four

hour session is the data basis. During the Google Glass session reaction was done by the speech command “bar changed” and on the consumer Tablet-PC reaction was done by double tapping on the screen. Furthermore on these two sessions 25 minutes performance was the data basis.

Results on reaction time show a significant effect in the industrial HMD study [$F(1, 40) = 5.409$, $p = .025$] indicating better reaction times on the industrial Tablet-PC, but no significant effect in the Google Glass study [$F(1, 34) = 1.850$, $p = .183$]. Reaction to colour change was significant faster in both studies [industrial: $F(1, 40) = 23.142$, $p < .001$; consumer: $F(1, 34) = 4.247$, $p = .047$] and the consumer HMD study also shows an expectable significant shorter reaction time for the single task baseline compared to the dual task [$F(1, 34) = 32.364$, $p < .001$].

Direct Comparison of subjective strain with both used HMDs

For subjective strain and visual fatigue parameters a group of participants (N=16) is used who experienced both HMDs. This is done to encounter individual answering tendencies which often overlap subjective ratings. The compared conditions are the second session with MAVUS HMD, the session with industrial Tablet-PC and the session with Google Glass in the other study.

Results of NASA-TLX show no significant difference this time [$F(2, 14) = .961$, $p = .406$]. In the original studies the HMDs had significant higher scores than the Tablet-PCs and failing of a significant effect might be based on limited sample size. Results of RSME show a significant effect of display [$F(2, 13) = 10.866$, $p = .002$] with highest values for the MAVUS HMD, lower values for Google Glass and lowest values for the Tabet-PC. Furthermore the increase over time becomes significant [$F(3, 12) = 8.254$, $p = .003$], but no interdependency display x time.

Values of the VFQ items are on a low overall level (0-3 on a 10 point scale) but show in many cases significant higher values for HMDs. For the items “difficulties to see sharp” [$F(2, 14) = 6.668$, $p = .009$] and “irritated / burning eyes” [$F(2, 14) = 3.458$, $p = .060$] the values for Google Glass are even higher after two hours of continuous use than on the MAVUS HMD, while values for Tablet-PC are near zero. The increase over time get also significant for both items [$F(2, 14) = 10.531$, $p = .002$ and $F(2, 14) = 6.350$, $p = .011$]. However, headache [$F(2, 13) = 7.003$, $p = .009$] and neck pain [$F(2, 13) = 4.357$, $p = .036$] have significantly the highest values for the MAVUS HMD with the heavier head carrier (means about 2 on a 10 point scale) after two hours while they stay close to zero for Google Glass and Tablet-PC. The general increase over time gets also significant [headache: $F(2, 13) = 5.153$, $p = .022$; neck pain: $F(2, 13) = 4.637$, $p = .030$] and also an interdependency display*time [headache: $F(4, 11) = 4.897$, $p = .016$; neck pain: $F(4, 11) = 3.394$, $p = .049$] indicating higher increase over time especially with the MAVUS HMD. The item “mental fatigue” also shows significant differences regarding display [$F(2, 13) = 6.522$, $p = .011$], an increase over time [$F(2, 13) = 8.831$, $p = .004$] and an interdependency display*time [$F(4, 11) = 4.447$, $p = .022$] indicating higher increase over time for both HMDs. While interpreting the alpha one has to keep in mind that the VFQ has no sum score and therefore theoretically the critical alpha has to be minimized by diverting .05 with the number of items

(16), so the new test value for significant effects will be .003 which will make significant effects very unlikely. On the other hand one might argue that the items of the VFQ do not necessarily represent the same phenomena as “difficulties to see sharp” and “headache” can also be independent from each other.

Discussion

In this paper we showed that performance in an assembling task does not profit from the new and lighter HMD Google Glass. It is even worse. And also reaction to a monitoring task is worse with the new HMD, as hit rate is now significantly below the Tablet-PC. One reason for this decrement in performance might be the monocular decoding of information on HMD, while on Tablet-PC both eyes can be used. Another possible reason could be that the position of the relevant information is more peripheral on HMD, while positioning of the Tablet-PC is free to the user. As Google Glass has a more peripheral position as MAVUS and performance is worse regarding assembling and monitoring task, this could be a hint that the positioning is crucial. Furthermore the “see-through” display in Google Glass could be irritating as subjects see the background slightly through, while MAVUS is a “look-around” display with no background coming through. However, in the study participants worked in front of a white wall, so background should not irritate them.

The worse reaction to monitoring task for HMD is true for parallel monitoring while other information is also presented on that display. But it does not say anything about pop-up alarms on a blank screen, which would be a complete different setup. However, for research projects that use an HMD mainly to display critical information within the field of view, hoping that therefore subjects will react more accurate to it, our findings are a strong hint to review this thesis.

The direct comparison of both studies has its limitations. Although the same combination of task was used, they differ in length and also in some other circumstances that might influence ratings too. But this rare occasion to compare two different HMDs is worth having a look at it. Although Google Glass has alike value in items regarding seeing sharp and burning eyes it has significant less values in headache and neck pain which makes this HMD much more comfortable than the industrial MAVUS HMD. This is in line with the answers to an interview question at the end of the study, where all participants prefer Google Glass to the MAVUS HMD. But in the end we have to say: Google Glass is more comfortable but not better in performance.

One question remains – as often when experimenting with new technology – to what amount effects will change or vanish with habituation? As HMDs are new and none of the subjects was used to work with them, it is quite logical that some decrements in performance or higher strain ratings are based on that fact rather than on the technology itself. In the MAVUS study we compared the first and second session with HMD: The performance stayed the same and also the subjective strain ratings, but the objective strain parameters of heart rate and heart rate variability showed on the second HMD session comparable strain as with the Tablet-PC. And habituation also might take longer than a four hour session. This indicates that we need more studies experimenting with the use of HMDs not only for half an hour, but for

prolonged time and even better for weeks or months. As HMDs will find their way into work places this should be accompanied by further studies in real situations and over longer periods.

References

Bangor, A.W. (2000). *Display technology and ambient illumination influences on visual fatigue at VDT Workstations.* PhD thesis, Virginia Polytechnic Institute and State University, USA. Online available: http://scholar.lib.vt.edu/theses/available/etd-03072001-091123/unrestricted/vfatigue.pdf [January 2015].

Grauel, B., Adolph, L. & Kluge, A. (2014). Head-Mounted Displays zur Unterstützung der örtlich getrennten Störungsdiagnose - passt die Technologie zur Aufgabe?. In Gesellschaft für Arbeitswissenschaft e. V., Gestaltung der Arbeitswelt der Zukunft, 60. Kongress der Gesellschaft für Arbeitswissenschaft. Dortmund: GfA-Press, pp. 691-693.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.) *Human mental workload* (pp. 139-183). Amsterdam: North-Holland.

Theis, S., Alexander, T., Mertens, A., Wille, M. & Schlick, C. (2014). Younger beginners, older retirees: Head-Mounted Displays and Demographic Change. Applied Human Factors and Ergonomics (AHFE), 19-23 June 2014 in Cracow, Poland.

Theis, S., Alexander, T., Mayer, M. & Wille, M. (2013). Considering ergonomic aspects of head-mounted displays for applications in industrial manufacturing. In: Duffy, V.G. (Ed.), Digital Human Modeling and Applications in Health, Safety, Ergonomics, and Risk Management (DHM/HCII) 2013, Part II, Lecture Notes in Computer Science 8026 (pp. 282-291). Berlin: Springer.

Wille, M., Grauel, B. & Adolph, L. (2013). Strain caused by head mounted displays. In: D. De Waard, , K. Brookhuis, R. Wiczorek, F. Di Nocera, P. Barham, C. Weikert, A. Kluge, W. Gerbino, and A. Toffetti, (Eds.) Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2013 Annual Conference (pp 267-277). HFES. Online available: http://www.hfes-europe.org/wp-content/uploads/2014/06/Wille.pdf [January 2015].

Wille, M., Scholl, P., Wischniewski, S. & Van Laerhoven, K. (2014). Comparing Google Glass with Tablet-PC as Guidance System for Assembling Tasks. Body Sensor Network, Zürich.

Wille, M., Wischniewski, S., Adolph, L., Theis, S., Grauel, B., & Alexander, T. (2014). Prolonged work with head mounted displays. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers: Adjunct Program* (pp. 221-224). ACM.

Zijlstra, F.R.H. (1993). *Efficiency in work behaviour: an approach for modern tools.* PhD thesis, University of Delft. Online available: http://resolver.tudelft.nl/uuid:d97a028b-c3dc-4930-b2ab-a7877993a17f [January 2015.